RESEARCH CENTRE

Inria Centre at the University of Lille

IN PARTNERSHIP WITH:

CNRS, Université de Lille

2024 ACTIVITY REPORT

Project-Team SCOOL

Sequential decision making under uncertainty problem

IN COLLABORATION WITH: Centre de Recherche en Informatique, Signal et Automatique de Lille

DOMAIN

Applied Mathematics, Computation and Simulation

THEME

Optimization, machine learning and statistical methods

Inria

Contents

Project-Team SCOOL

Creation of the Project-Team: 2020 November 01

Keywords

Computer sciences and digital sciences

- A3. – Data and knowledge
- A3.1. – Data
- A3.1.1. – Modeling, representation
- A3.1.4. – Uncertain data
- A3.1.11. – Structured data
- A3.3. – Data and knowledge analysis
- A3.3.1. – On-line analytical processing
- A3.3.2. – Data mining
- A3.3.3. – Big data analysis
- A3.4. – Machine learning and statistics
- A3.4.1. – Supervised learning
- A3.4.2. – Unsupervised learning
- A3.4.3. – Reinforcement learning
- A3.4.4. – Optimization and learning
- A3.4.5. – Bayesian methods
- A3.4.6. – Neural networks
- A3.4.8. – Deep learning
- A3.5.2. – Recommendation systems
- A5.1. – Human-Computer Interaction
- A5.10.7. – Learning
- A8.6. – Information theory
- A8.11. – Game Theory
- A9. – Artificial intelligence
- A9.2. – Machine learning
- A9.3. – Signal analysis
- A9.4. – Natural language processing
- A9.7. – AI algorithmics

Other research topics and application domains

- B2. – Health
- B3.1. – Sustainable development
- B3.5. – Agronomy
- B9.5. – Sciences
- B9.5.6. – Data science

1 Team members, visitors, external collaborators

Research Scientists

- Riadh Akrour [INRIA, ISFP]
- Debabrota Basu [INRIA, ISFP]
- Remy Degenne [INRIA, ISFP]
- Emilie Kaufmann [CNRS, Researcher]
- Odalric-Ambrym Maillard [INRIA, Researcher]
- Timothée Mathieu [INRIA, Researcher]

Faculty Members

- Philippe Preux [Team leader, UNIV LILLE, Professor Delegation]
- Juliette Achddou [UNIV LILLE, Associate Professor, from Sep 2024]

Post-Doctoral Fellows

- Sabrine Chebbi [INRIA, Post-Doctoral Fellow, from Sep 2024]
- Tuan Dam Quang Tuan [INRIA, Post-Doctoral Fellow, until Sep 2024]
- Tanguy Lefort [INRIA, Post-Doctoral Fellow, from Oct 2024]
- Alena Shilova [INRIA]

PhD Students

- Ayoub Ajarra [INRIA]
- Achraf Azize [UNIV LILLE]
- Mickael Basson [LILLY FRANCE, CIFRE]
- Yann Berthelot [Saint-Gobain Research, CIFRE]
- Udvas Das [INRIA, from Mar 2024]
- Brahim Driss [INRIA, from Oct 2024]
- Marc Jourdan [UNIV LILLE, until Sep 2024]
- Anthony Kobanda [UBISOFT]
- Hector Kohler [UNIV LILLE]
- Penanklihi Cyrille Kone [INRIA]
- Matheus Medeiros Centa [UNIV LILLE]
- Thomas Meunier [INRIA, until Sep 2024]
- Thomas Michel [INRIA, from Oct 2024]
- Adrien Prevost [INRIA, from Nov 2024]
- Waris Radji [INRIA, from Oct 2024]
- Adrienne Tuynman [ENS PARIS-SACLAY]
- Sumit Vashishtha [UNIV LILLE]

Technical Staff

- Hernan David Carvajal Bastidas [INRIA, Engineer, until Aug 2024]
- Alex Davey [INRIA, Engineer, from May 2024]
- Brahim Driss [INRIA, Engineer, until Sep 2024]
- Guillaume Pourcel [INRIA, Engineer, from Feb 2024 until Apr 2024]
- Waris Radji [INRIA, Engineer, until Aug 2024]
- Julien Teigny [INRIA, Engineer]

Interns and Apprentices

- Corentin Blondelle [UNIV LILLE, Intern, from Apr 2024 until Jun 2024]
- Mohamed Yassine Kabouri [INRIA, Intern, from Apr 2024 until Sep 2024]
- Lorenzo Luccioli [INRIA, Intern, from Apr 2024 until Sep 2024]
- Gabriele Maggioni [INRIA, Intern, from May 2024 until Sep 2024]
- Antoine Olivier [Bits2beat, Intern, until Nov 2024]
- Riccardo Poiani [ECOLE POLYT. MILAN, Intern, from Mar 2024 until Jun 2024]
- Adrien Prevost [INRIA, Intern, from Apr 2024 until Sep 2024]
- Quentin Uguen [CENTRALE59, Intern, from Jul 2024 until Aug 2024]

Administrative Assistants

- Aurore Dalle [INRIA]
- Lucille Leclercq [INRIA]
- Anne Rejl [INRIA]
- Amélie Supervielle [INRIA]

Visiting Scientist

- Ayman Chaouki [Ecole Polytechnique, from Sep 2024 until Sep 2024]

External Collaborators

- Riccardo Della Vecchia [HSBC, from Jun 2024 until Nov 2024]
- Fabien Pesquerel [UNIV LILLE, until Feb 2024]

2 Overall objectives

Scool is a machine learning (ML) research group. Scool's research focuses on the study of the sequential decision making under uncertainty problem (SDMUP). In particular, we consider bandit problems [46] and the reinforcement learning (RL) problem [45]. In a simplified way, RL considers the problem of learning an optimal policy in a Markov Decision Problem (MDP) [43]; when the set of states collapses to a single state, this is known as the bandit problem which focuses on the exploration/exploitation problem.

Bandit and RL problems are interesting to study on their own; both types of problems share a number of fundamental issues (convergence analysis, sample complexity, representation, safety, *etc.*); both problems have real life applications, different though closely related; the fact that while solving an RL problem, one faces an exploration/exploitation problem and has to solve a bandit problem in each state connects the two types of problems very intimately.

In our work, we also consider settings going beyond the Markovian assumption, in particular non-stationary settings, which represent a challenge common to bandits and RL. A distinctive aspect of the SDMUP with regards to the rest of the field of ML is that the learning problem takes place within a closed-loop interaction between a learning agent and its environment. This feedback loop makes our field of research very different from the two other sub-fields of ML, supervised and unsupervised learning, even when they are defined in an incremental setting. Hence, SDMUP combines ML with control: the learner is not passive, the learner acts on its environment, and learns from the consequences of these interactions; hence, the learner can act in order to obtain information from the environment. Naturally, the optimal control community is getting more and more interested by RL (see e.g. [44]).

We wish to go on, studying applied questions and developing theory to come up with sound approaches to the practical resolution of SDMUP tasks, and guide their resolution. Non-stationary environments are a particularly interesting setting; we are studying this setting and developing new tools to approach it in a sound way, in order to have algorithms to detect environment changes as fast as possible, and as reliably as possible, adapt to them, and prove their behavior, in terms of their performance, measured with the regret for instance. We mostly consider non parametric statistical models, that is models in which the number of parameters is not fixed (a parameter may be of any type: a scalar, a vector, a function, *etc.*), so that the model can adapt along learning, and to its changing environment; this also lets the algorithm learn a representation that fits its environment.

3 Research program

Our research is mostly dealing with bandit problems, and reinforcement learning problems. We investigate each thread separately and also in combination, since the management of the exploration/exploitation trade-off is a major issue in reinforcement learning.

On bandit problems, we focus on:

- structured bandits
- bandits for planning (in particular for Monte Carlo Tree Search (MCTS))
- non stationary bandits

Regarding reinforcement learning, we focus on:

- modeling issues, and dealing with the discrepancy between the model and the task to solve
- learning and using the structure of a Markov decision problem, and of the learned policy
- generalization in reinforcement learning
- reinforcement learning in non stationary environments

Beyond these objectives, we put a particular emphasis on the study of non-stationary environments. Another area of great concern is the combination of symbolic methods with numerical methods, be it to

provide knowledge to the learning algorithm to improve its learning curve, or to better understand what the algorithm has learned and explain its behavior, or to rely on causality rather than on mere correlation.

We also put a particular emphasis on real applications and how to deal with their constraints: lack of a simulator, difficulty to have a realistic model of the problem, small amount of data, dealing with risks, availability of expert knowledge on the task.

4 Application domains

Scool has 2 main topics of application:

- health
- sustainable development

In each of these two domains, we put forward the investigation and the application of the idea of sequential decision making under uncertainty. Though supervised and non supervised learning have already been studied and applied extensively, sequential decision making remains far less studied; bandits have already been used in many applications of e-commerce (e.g. for computational advertising and recommendation systems). However, in applications where human beings may be severely impacted, bandits and reinforcement learning have not been studied much; moreover, these applications come along with a scarcity of data, and the non availability of a simulator, which prevents heavy computational simulations to come up with safe automatic decision making.

In 2022, in health, we investigated patient follow-up with Prof. F. Pattou's research group (CHU Lille, Inserm, Université de Lille) in project B4H. This effort came along with investigating how we may use medical data available locally at CHU Lille, and also the national social security data. We also investigated drug repurposing with Prof. A. Delahaye-Duriez (Inserm, Université de Paris) in project Repos. We also studied catheter control by way of reinforcement learning with Inria Lille group Defrost, and company Robocath (Rouen).

Regarding sustainable development, we have a set of projects and collaborations regarding agriculture and gardening. With Cirad and CGIAR, we investigate how one may recommend agricultural practices to farmers in developing countries. Through an associate team with Bihar Agriculture University (India), we investigate data collection. Inria exploratory action SR4SG concerns recommender systems at the level of individual gardens.

There are two important aspects that are amply shared by these two application fields. First, we consider that data collection is an active task: we do not passively observe and record data, we design methods and algorithms to search for useful data. This idea is exploited in most of these works oriented towards applications. Second, many of these projects include a careful management of risks for human beings. We have to take decisions taking care of their consequences on human beings, on eco-systems and life more generally.

5 Social and environmental responsibility

Sustainable development is a major field of research and application of Scool. We investigate what machine learning can bring to sustainable development, identifiying challenges and obstacles, and studying how to overcome them.

Let us mention here:

- sustainable agriculture in developing countries;
- sustainable gardening.

More details can be found in Section 4.

6 Highlights of the year

6.1 Awards

Emilie Kaufmann was awarded the CNRS bronze medal.

7 New software, platforms, open data

7.1 New software

7.1.1 rlberry

Keywords: Reinforcement learning, Simulation, Artificial intelligence

Scientific Description: RL-berry is a software library for reinforcement learning. One of its major features is to give access to state-of-the-art RL algorithm implementations in a very simple way. One of the goal is that RL-berry would be used for teaching. We also provide original features that are missing from other RL libraries. One concerns the methodology to compare the experimental performance of various algorithms.

Functional Description: rlberry is a reinforcement learning (RL) library in Python for research and education. The library provides implementations of several RL agents for you to use as a starting point or as baselines, provides a set of benchmark environments, very useful to debug and challenge your algorithms, handles all random seeds for you, ensuring reproducibility of your results, and is fully compatible with several commonly used RL libraries like OpenAI gym and Stable Baselines.

URL: https://github.com/rlberry-py/rlberry

Contact: Timothee Mathieu

7.1.2 Weight Trajectory Predictor : algorithm

Name: Weight Trajectory Predictor : algorithm

Keywords: Medical applications, Machine learning

Scientific Description: We performed a retrospective study of clinical data collected prospectively on patients with up to five years postoperative follow-up (ABOS cohort, CHU Lille) and trained a supervised model to predict the relative total weight loss ("%TWL") of a patient 1, 3, 12, 24 and 60 months after surgery. This model consists in a decision tree, written in python, taking as input a selected subset of preoperative attributes (weight, height, type of intervention, age, presence or absence of type 2 diabetes or impaired glucose tolerance, diabetes duration, smoking habits) and returns an estimation of %TWL as well as a prediction interval based on the interquartile range of %TWL observed on similar patients. The predictions of this tool have been validated both internally and externally (on French and Dutch cohorts).

Functional Description: The "Weight Tracjectory Predictor" algorithm is part of a larger project, whose goal is to leverage artificial intelligence techniques to improve patient care. This code is the product of a collaboration between Inria SCOOL and the UMR 1190-EGID team of the CHU Lille. It aims to predict the weight loss trajectory of a patient following bariatric surgery (treatment of severe obesity) from a set of preoperative characteristics.

We performed a retrospective study of clinical data collected prospectively on patients with up to five years postoperative follow-up (ABOS cohort, CHU Lille) and trained a supervised model to predict the relative total weight loss ("%TWL") of a patient 1, 3, 12, 24 and 60 months after surgery. This model consists in a decision tree, written in python, taking as input a selected subset of preoperative attributes (weight, height, type of intervention, age, presence or absence of type 2 diabetes or impaired glucose tolerance, diabetes duration, smoking habits) and returns an estimation of %TWL as well as a prediction interval based on the interquartile range of %TWL

observed on similar patients. The predictions of this tool have been validated both internally and externally (on French and Dutch cohorts).

The goal of this software is to improve patient follow-up after bariatric surgery: - during preoperative visits, by providing clinicians with a quantitative tool to inform the patient regarding potential weight loss outcome. - during postoperative control visits, by comparing the predicted and realized weight trajectories, which may facilitate early detection of complications.

This software component will be embedded in a web app for ease of use.

Release Contributions: Initial version

URL: https://bariatric-weight-trajectory-prediction.univ-lille.fr/

Contact: Julien Teigny

Participants: Pierre Bauvin, Francois Pattou, Philippe Preux, Violeta Raverdy, Patrick Saux, Tomy Soumphonphakdy, Julien Teigny, Hélène Verkindt

Partner: CHU de Lille

7.1.3 Adastop

Keywords: Hypothesis testing, Reinforcement learning, Reproducibility

Functional Description: This package contains the AdaStop algorithm. AdaStop implements a statistical test to adaptively choose the number of runs of stochastic algorithms necessary to compare these algorithms and be able to rank them with a theoretically controlled family-wise error rate. One particular application for which AdaStop was created is to compare Reinforcement Learning algorithms. Please note, that what we call here an algorithm is really a certain implementation of an algorithm.

URL: https://github.com/TimotheeMathieu/adastop

Contact: Timothee Mathieu

7.1.4 average-reward-reinforcement-learning

Keywords: Mutli-armed bandits, Reinforcement learning, Python

Functional Description: Library of RL and Bandit algorithms.

URL: https://gitlab.inria.fr/omaillar/average-reward-reinforcement-learning

Contact: Odalric-Ambrym Maillard

Participant: Odalric-Ambrym Maillard

7.1.5 FarmGym

Name: Farming Environment Gym factory for Reinforcement Learning

Keywords: Reinforcement learning, Simulator, Agroecology

Functional Description: Farming Environment Gym factory for Reinforcement Learning

Release Contributions: This version is an entire rewriting by Odlaric-Ambrym Maillard of the prototype V1 created by Thomas Carta. Version V2 features modular creation of farms, specifications of various entities and monitoring facilities. Recent additions include slight adjustment of the entities dynamics. It also features unit tests and an automatic generation of base policies done by Brahim Driss.

News of the Year: Nov 2022: Déploiement de FarmGym sur gitlab, auparavant en développement interne par Odalric-Ambrym Maillard (2021-2022). Nov-Dec 2022: Organisation de compétition interne par Timothée Mathieu.

Printemps 2023: Arrivée de Brahim Driss sur le projet, pour mettre en place tests unitaires/fonctionnels et assister Odalric-Ambrym Maillard dans le développement de fonctionnalité.

URL: https://github.com/farm-gym/

Publication: hal-03960683

Contact: Odalric-Ambrym Maillard

Participants: Odalric-Ambrym Maillard, Brahim Driss, Timothee Mathieu

Partner: Inria

7.1.6 Dyjest algorithms

Name: Dyjest Algorithms Toolbox

Keywords: NLP, Statistic analysis

Scientific Description: Smart Tracker uses a combination of Named Entity Recognition (NER) and Language Model Recognition (LLM) methods to structure user input. This data is then transformed into embeddings and processed by semantic search algorithms to find the closest matches in a database of symptoms, foods and activities. Finally, a report generator that presents the results of statistical methods for identifying potential symptom triggers.

Functional Description: This code repository includes work carried out on the following aspects: - Streamlining the tracking of meals, activities, and symptoms through the entry of unstructured textual data. - Using embeddings with fine-tuning to improve the matching of inputs. - Integrating a hierarchical structure (tree) to enhance the robustness of semantic search. - Generating personalized reports. - Exploring symptom prediction (inconclusive) and applying multi-armed bandits to food recommendations.

Contact: Medhi Douch

Participants: Hatim Mrabet, Naafi Dasana Ibrahim, Emilie Kaufmann, Medhi Douch

8 New results

We organize our research results in a set of categories. The main categories are: bandits and RL theory, bandits and RL under real life constraints, and applications.

Participants: all Scool members.

8.1 Bandits and RL theory

Finding good policies in average-reward Markov Decision Processes without prior knowledge, [28]

We revisit the identification of an ε-optimal policy in average-reward Markov Decision Processes (MDP). In such MDPs, two measures of complexity have appeared in the literature: the diameter, D, and the optimal bias span, H, which satisfy $H \leq D$. Prior work have studied the complexity of ε-optimal policy identification only when a generative model is available. In this case, it is known that there exists an MDP with $D \simeq H$ for which the sample complexity to output an ε-optimal policy is $\Omega(SAD/\varepsilon^2)$ where S and A are the sizes of the state and action spaces. Recently, an algorithm with a sample complexity of order

SAH/ε^2 has been proposed, but it requires the knowledge of H. We first show that the sample complexity required to estimate H is not bounded by any function of S, A and H, ruling out the possibility to easily make the previous algorithm agnostic to H. By relying instead on a diameter estimation procedure, we propose the first algorithm for (ε, δ)-PAC policy identification that does not need any form of prior knowledge on the MDP. Its sample complexity scales in SAD/ε^2 in the regime of small ε, which is near-optimal. In the online setting, our first contribution is a lower bound which implies that a sample complexity polynomial in H cannot be achieved in this setting. Then, we propose an online algorithm with a sample complexity in SAD^2/ε^2, as well as a novel approach based on a data-dependent stopping rule that we believe is promising to further reduce this bound.

Power Mean Estimation in Stochastic Monte-Carlo Tree Search, [22]

Monte-Carlo Tree Search (MCTS) is a widely-used strategy for online planning that combines Monte-Carlo sampling with forward tree search. Its success relies on the Upper Confidence bound for Trees (UCT) algorithm, an extension of the UCB method for multi-arm bandits. However, the theoretical foundation of UCT is incomplete due to an error in the logarithmic bonus term for action selection, leading to the development of Fixed-Depth-MCTS with a polynomial exploration bonus to balance exploration and exploitation. Both UCT and Fixed-Depth-MCTS suffer from biased value estimation: the weighted sum underestimates the optimal value, while the maximum valuation overestimates it. The power mean estimator offers a balanced solution, lying between the average and maximum values. Power-UCT (Dam et al. 2019) incorporates this estimator for more accurate value estimates but its theoretical analysis remains incomplete. This paper introduces Stochastic-Power-UCT, an MCTS algorithm using the power mean estimator and tailored for stochastic MDPs. We analyze its polynomial convergence in estimating root node values and show that it shares the same convergence rate as Fixed-Depth-MCTS, with the latter being a special case of the former. Our theoretical results are validated with empirical tests across various stochastic MDP environments.

Bandits with Multimodal Structure, [26]

We consider a multi-armed bandit problem specified by a set of one-dimensional exponential family distributions endowed with a multimodal structure. The multimodal structure naturally extends the unimodal structure and appears to be underlying in quite interesting ways popular structures such as linear or Lipschitz bandits. We introduce IMED-MB, an algorithm that optimally exploits the multimodal structure, by adapting to this setting the popular Indexed Minimum Empirical Divergence (IMED) algorithm. We provide instance-dependent regret analysis of this strategy. Numerical experiments show that IMED-MB performs well in practice when assuming unimodal, polynomial or Lipschitz mean function.

Preference-based Pure Exploration, [27]

We study the preference-based pure exploration problem for bandits with vector-valued rewards ordered using a preference cone $\mathscr{C}$ with the goal of identifying the most preferred policy over the set of arms. First, to quantify the impact of preferences, we derive a novel lower bound on the sample complexity for identifying the most preferred policy with confidence level $1-\delta$. Our lower bound elicits the role played by the geometry of the preference cone and punctuates the difference in hardness compared to best-arm variants of the problem. We further explicate this geometry when rewards follow a Gaussian distributions, and provide a convex reformulation of the lower bound. Then, we leverage this convex reformulation of the lower bound to design the Preference-based Track and Stop (PreTS) algorithm that identifies the most preferred policy. Finally, we derive a new concentration result for vector-valued rewards, and show that PreTS achieves a matching sample complexity upper bound.

8.2 Bandits and RL under Real-life constraints

8.2.1 Linear constraints

Pure Exploration in Bandits with Linear Constraints, [21]

We address the problem of identifying the optimal policy with a fixed confidence level in a multi-armed bandit setup, when the arms are subject to linear constraints. Unlike the standard best-arm identification problem which is well studied, the optimal policy in this case may not be deterministic and could mix between several arms. This changes the geometry of the problem which we characterize via an information-theoretic lower bound. We introduce two asymptotically optimal algorithms for this setting,

one based on the Track-and-Stop method and the other based on a game-theoretic approach. Both these algorithms try to track an optimal allocation based on the lower bound and computed by a weighted projection onto the boundary of a normal cone. Finally, we provide empirical results that validate our bounds and visualize how constraints change the hardness of the problem.

Learning to Explore with Lagrangians for Bandits under Unknown Linear Constraints, [30]

Pure exploration in bandits models multiple real-world problems, such as tuning hyper-parameters or conducting user studies, where different safety, resource, and fairness constraints on the decision space naturally appear. We study these problems as pure exploration in multi-armed bandits with unknown linear constraints, where the aim is to identify an *r-good feasible policy*. First, we propose a Lagrangian relaxation of the sample complexity lower bound for pure exploration under constraints. We show how this lower bound evolves with the sequential estimation of constraints. Second, we leverage the Lagrangian lower bound and the properties of convex optimisation to propose two computationally efficient extensions of Track-and-Stop and Gamified Explorer, namely LATS and LAGEX. To this end, we propose a constraint-adaptive stopping rule, and while tracking the lower bound, use pessimistic estimate of the feasible set at each step. We show that these algorithms achieve asymptotically optimal sample complexity upper bounds up to constraint-dependent constants. Finally, we conduct numerical experiments with different reward distributions and constraints that validate efficient performance of LAGEX and LATS with respect to baselines.

8.2.2 Robustness

CRIMED: Lower and Upper Bounds on Regret for Bandits with Unbounded Stochastic Corruption, [18]

We investigate the regret-minimisation problem in a multi-armed bandit setting with arbitrary corruptions. Similar to the classical setup, the agent receives rewards generated independently from the distribution of the arm chosen at each time. However, these rewards are not directly observed. Instead, with a fixed $\varepsilon \in (0, \frac{1}{2})$, the agent observes a sample from the chosen arm's distribution with probability $1-\varepsilon$, or from an arbitrary corruption distribution with probability ε. Importantly, we impose no assumptions on these corruption distributions, which can be unbounded. In this setting, accommodating potentially unbounded corruptions, we establish a problem-dependent lower bound on regret for a given family of arm distributions. We introduce CRIMED, an asymptotically-optimal algorithm that achieves the exact lower bound on regret for bandits with Gaussian distributions with known variance. Additionally, we provide a finite-sample analysis of CRIMED's regret performance. Notably, CRIMED can effectively handle corruptions with ε values as high as $\frac{1}{2}$. Furthermore, we develop a tight concentration result for medians in the presence of arbitrary corruptions, even with ε values up to $\frac{1}{2}$, which may be of independent interest. We also discuss an extension of the algorithm for handling misspecification in Gaussian model.

Bandits Corrupted by Nature: Lower Bounds on Regret and Robust Optimistic Algorithms, [14]

We study the Bandits with Stochastic Corruption problem, i.e. a stochastic multi-armed bandit problem with k unknown reward distributions, which are heavy-tailed and corrupted by a history-independent stochastic adversary or Nature. To be specific, the reward obtained by playing an arm comes from corresponding heavy-tailed reward distribution with probability $1-\varepsilon \in (0.5, 1]$ and an arbitrary corruption distribution of unbounded support with probability $\varepsilon \in [0, 0.5)$. First, we provide *a problem-dependent lower bound on the regret* of any corrupted bandit algorithm. The lower bounds indicate that the Bandits with Stochastic Corruption problem is harder than the classical stochastic bandit problem with sub-Gaussian or heavy-tail rewards. Following that, we propose a novel UCB-type algorithm for Bandits with Stochastic Corruption, namely HuberUCB, that builds on Huber's estimator for robust mean estimation. Leveraging a novel concentration inequality of Huber's estimator, we prove that HuberUCB achieves a near-optimal regret upper bound. Since computing Huber's estimator has quadratic complexity, we further introduce a sequential version of Huber's estimator that exhibits linear complexity. We leverage this sequential estimator to design SequentialHuberUCB that enjoys similar regret guarantees while reducing the computational burden. Finally, we experimentally illustrate the efficiency of HuberUCB and SequentialHuberUCB in solving Bandits with Stochastic Corruption for different reward distributions and different levels of corruptions.

8.2.3 Privacy

Concentrated Differential Privacy for Bandits, [19]

Bandits serve as the theoretical foundation of sequential learning and an algorithmic foundation of modern recommender systems. However, recommender systems often rely on user-sensitive data, making privacy a critical concern. This paper contributes to the understanding of Differential Privacy (DP) in bandits with a trusted centralised decision-maker, and especially the implications of ensuring zero Concentrated Differential Privacy (zCDP). First, we formalise and compare different adaptations of DP to bandits, depending on the considered input and the interaction protocol. Then, we propose three private algorithms, namely AdaC-UCB, AdaC-GOPE and AdaC-OFUL, for three bandit settings, namely finite-armed bandits, linear bandits, and linear contextual bandits. The three algorithms share a generic algorithmic blueprint, i.e. the Gaussian mechanism and adaptive episodes, to ensure a good privacy-utility trade-off. We analyse and upper bound the regret of these three algorithms. Our analysis shows that in all of these settings, the prices of imposing zCDP are (asymptotically) negligible in comparison with the regrets incurred oblivious to privacy. Next, we complement our regret upper bounds with the first minimax lower bounds on the regret of bandits with zCDP. To prove the lower bounds, we elaborate a new proof technique based on couplings and optimal transport. We conclude by experimentally validating our theoretical results for the three different settings of bandits.

FLIPHAT: Joint Differential Privacy for High Dimensional Sparse Linear Bandits, [38]

High dimensional sparse linear bandits serve as an efficient model for sequential decision-making problems (e.g. personalized medicine), where high dimensional features (e.g. genomic data) on the users are available, but only a small subset of them are relevant. Motivated by data privacy concerns in these applications, we study the joint differentially private high dimensional sparse linear bandits, where both rewards and contexts are considered as private data. First, to quantify the cost of privacy, we derive a lower bound on the regret achievable in this setting. To further address the problem, we design a computationally efficient bandit algorithm, **F**orgetfu**L** **I**terative **P**rivate **HA**rd **T**hresholding (FLIPHAT). Along with doubling of episodes and episodic forgetting, FLIPHAT deploys a variant of Noisy Iterative Hard Thresholding (N-IHT) algorithm as a sparse linear regression oracle to ensure both privacy and regret-optimality. We show that FLIPHAT achieves optimal regret up to logarithmic factors. We analyze the regret by providing a novel refined analysis of the estimation error of N-IHT, which is of parallel interest.

Differentially Private Best-Arm Identification, [35]

Best Arm Identification (BAI) problems are progressively used for data-sensitive applications, such as designing adaptive clinical trials, tuning hyper-parameters, and conducting user studies. Motivated by the data privacy concerns invoked by these applications, we study the problem of BAI with fixed confidence in both the local and central models, i.e. ϵ-local and ϵ-global Differential Privacy (DP). First, to quantify the cost of privacy, we derive lower bounds on the sample complexity of any δ-correct BAI algorithm satisfying ϵ-global DP or ϵ-local DP. Our lower bounds suggest the existence of two privacy regimes. In the high-privacy regime, the hardness depends on a coupled effect of privacy and novel information-theoretic quantities involving the Total Variation. In the low-privacy regime, the lower bounds reduce to the non-private lower bounds. We propose ϵ-local DP and ϵ-global DP variants of a Top Two algorithm, namely CTB-TT and AdaP-TT*, respectively. For ϵ-local DP, CTB-TT is asymptotically optimal by plugging in a private estimator of the means based on Randomised Response. For ϵ-global DP, our private estimator of the mean runs in arm-dependent adaptive episodes and adds Laplace noise to ensure a good privacy-utility trade-off. By adapting the transportation costs, the expected sample complexity of AdaP-TT* reaches the asymptotic lower bound up to multiplicative constants.

Open Problem: What is the Complexity of Joint Differential Privacy in Linear Contextual Bandits?, [20]

Contextual bandits serve as a theoretical framework to design recommender systems, which often rely on user-sensitive data, making privacy a critical concern. However, a significant gap remains between the known upper and lower bounds on the regret achievable in linear contextual bandits under Joint Differential Privacy (JDP), which is a popular privacy definition used in this setting. In particular, the best regret upper bound is known to be $O\left(d\sqrt{T}\log(T) + d^{3/4}\sqrt{T\log(1/\delta)}/\sqrt{\epsilon}\right)$, while the lower bound is $\Omega\left(\sqrt{dT\log(K)} + d/(\epsilon+\delta)\right)$. We discuss the recent progress on this problem, both from the algorithm design and lower bound techniques, and posit the open questions.

8.2.4 Multi-fidelity feedback

Optimal Multi-Fidelity Best-Arm Identification, [25]

In bandit best-arm identification, an algorithm is tasked with finding the arm with highest mean reward with a specified accuracy as fast as possible. We study multi-fidelity best-arm identification, in which the algorithm can choose to sample an arm at a lower fidelity (less accurate mean estimate) for a lower cost. Several methods have been proposed for tackling this problem, but their optimality remain elusive, notably due to loose lower bounds on the total cost needed to identify the best arm. Our first contribution is a tight, instance-dependent lower bound on the cost complexity. The study of the optimization problem featured in the lower bound provides new insights to devise computationally efficient algorithms, and leads us to propose a gradient-based approach with asymptotically optimal cost complexity. We demonstrate the benefits of the new algorithm compared to existing methods in experiments. Our theoretical and empirical findings also shed light on an intriguing concept of optimal fidelity for each arm.

8.3 Bandits and RL for real-life: Deep RL and Applications

AdaStop: sequential testing for efficient and reliable comparisons of Deep RL Agents, [15]

Recently, the scientific community has questioned the statistical reproducibility of many empirical results, especially in the field of machine learning. To contribute to the resolution of this reproducibility crisis, we propose a theoretically sound methodology for comparing the performance of a set of algorithms. We exemplify our methodology in Deep Reinforcement Learning (Deep RL). The performance of one execution of a Deep RL algorithm is a random variable. Therefore, several independent executions are needed to evaluate its performance. When comparing algorithms with random performance, a major question concerns the number of executions to perform to ensure that the result of the comparison is theoretically sound. Researchers in Deep RL often use less than 5 independent executions to compare algorithms: we claim that this is not enough in general. Moreover, when comparing more than 2 algorithms at once, we have to use a multiple tests procedure to preserve low error guarantees. We introduce AdaStop, a new statistical test based on multiple group sequential tests. When used to compare algorithms, AdaStop adapts the number of executions to stop as early as possible while ensuring that enough information has been collected to distinguish algorithms that have different score distributions. We prove theoretically that AdaStop has a low probability of making a (family-wise) error. We illustrate the effectiveness of AdaStop in various use-cases, including toy examples and Deep RL algorithms on challenging Mujoco environments. AdaStop is the first statistical test fitted to this sort of comparisons: it is both a significant contribution to statistics, and an important contribution to computational studies performed in reinforcement learning and in other domains.

Interpretable and Editable Programmatic Tree Policies for Reinforcement Learning, [24]

Deep reinforcement learning agents are prone to goal misalignments. The black-box nature of their policies hinders the detection and correction of such misalignments, and the trust necessary for real-world deployment. So far, solutions learning interpretable policies are inefficient or require many human priors. We propose INTERPRETER, a fast distillation method producing INTerpretable Editable tRee Programs for ReinforcEmenT lEaRning. We empirically demonstrate that INTERPRETER compact tree programs match oracles across a diverse set of sequential decision tasks and evaluate the impact of our design choices on interpretability and performances. We show that our policies can be interpreted and edited to correct misalignments on Atari games and to explain real farming strategies.

Learning HJB Viscosity Solutions with PINNs for Continuous-Time Reinforcement Learning, [42]

Despite recent advances in Reinforcement Learning (RL), the Markov Decision Processes are not always the best choice to model complex dynamical systems requiring interactions at high frequency. Being able to work with arbitrary time intervals, Continuous Time Reinforcement Learning (CTRL) is more suitable for those problems. Instead of the Bellman equation operating in discrete time, it is the Hamilton-Jacobi-Bellman (HJB) equation that describes value function evolution in CTRL. Even though the value function is a solution of the HJB equation, it may not be its unique solution. To distinguish the value function from other solutions, it is important to look for the viscosity solutions of the HJB equation. The viscosity solutions constitute a special class of solutions that possess uniqueness and

stability properties. This paper proposes a novel approach to approximate the value function by training a physics informed neural network (PINN) through a specific ϵ-scheduling iterative process constraining the PINN to converge towards the viscosity solution and shows experimental results with classical control tasks, where PINNs outperform popular RL algorithms in a nearly continuous-time setting.

Augmented Bayesian Policy Search, [23]

Deterministic policies are often preferred over stochastic ones when implemented on physical systems. They can prevent erratic and harmful behaviors while being easier to implement and interpret. However, in practice, exploration is largely performed by stochastic policies. First-order Bayesian optimization (BO) methods offer a principled way of performing exploration using deterministic policies. This is done through a learned probabilistic model of the objective function and its gradient. Nonetheless, such approaches treat policy search as a black-box problem, and thus, neglect the reinforcement learning nature of the problem. In this work, we leverage the performance difference lemma to introduce a novel mean function for the probabilistic model. This results in augmenting BO methods with the action-value function. Hence, we call our method Augmented Bayesian Search (ABS). Interestingly, this new mean function enhances the posterior gradient with the deterministic policy gradient, effectively bridging the gap between BO and policy gradient methods. The resulting algorithm combines the convenience of the direct policy search with the scalability of reinforcement learning. We validate ABS on high-dimensional locomotion problems and demonstrate competitive performance compared to existing direct policy search schemes.

Reinforcement Learning in the Wild with Maximum Likelihood-based Model Transfer, [31]

In this paper, we study the problem of transferring the available Markov Decision Process (MDP) models to learn and plan efficiently in an unknown but similar MDP. We refer to it as *Model Transfer Reinforcement Learning (MTRL)* problem. First, we formulate MTRL for discrete MDPs and Linear Quadratic Regulators (LQRs) with continuous state actions. Then, we propose a generic two-stage algorithm, MLEMTRL, to address the MTRL problem in discrete and continuous settings. In the first stage, MLEMTRL uses a *constrained Maximum Likelihood Estimation (MLE)*-based approach to estimate the target MDP model using a set of known MDP models. In the second stage, using the estimated target MDP model, MLEMTRL deploys a model-based planning algorithm appropriate for the MDP class. Theoretically, we prove worst-case regret bounds for MLEMTRL both in realisable and non-realisable settings. We empirically demonstrate that MLEMTRL allows faster learning in new MDPs than learning from scratch and achieves near-optimal performance depending on the similarity of the available MDPs and the target MDP.

Dynamical-VAE-based Hindsight to Learn the Causal Dynamics of Factored-POMDPs, [40]

Learning representations of underlying environmental dynamics from partial observations is a critical challenge in machine learning. In the context of Partially Observable Markov Decision Processes (POMDPs), state representations are often inferred from the history of past observations and actions. We demonstrate that incorporating future information is essential to accurately capture causal dynamics and enhance state representations. To address this, we introduce a Dynamical Variational Auto-Encoder (DVAE) designed to learn causal Markovian dynamics from offline trajectories in a POMDP. Our method employs an extended hindsight framework that integrates past, current, and multi-step future information within a factored-POMDP setting. Empirical results reveal that this approach uncovers the causal graph governing hidden state transitions more effectively than history-based and typical hindsight-based models.

Measuring Exploration in Reinforcement Learning via Optimal Transport in Policy Space, [41]

Exploration is the key ingredient of reinforcement learning (RL) that determines the speed and success of learning. Here, we quantify and compare the amount of exploration and learning accomplished by a Reinforcement Learning (RL) algorithm. Specifically, we propose a novel measure, named Exploration Index, that quantifies the relative effort of knowledge transfer (transferability) by an RL algorithm in comparison to supervised learning (SL) that transforms the initial data distribution of RL to the corresponding final data distribution. The comparison is established by formulating learning in RL as a sequence of SL tasks, and using optimal transport based metrics to compare the total path traversed by the RL and SL algorithms in the data distribution space. We perform extensive empirical analysis on various environments and with multiple algorithms to demonstrate that the exploration index yields insights about the exploration behaviour of any RL algorithm, and also allows us to compare the exploratory

behaviours of different RL algorithms.

8.4 Others

8.4.1 Statistical reproductibility in practice

Statistical comparison in empirical computer science with minimal computation usage, [32]

The replicability of computational experiments remains a fundamental question. Very often, computational experiments are meant to compare the performance of different algorithms on a given task or on a given set of tasks. For example, the machine learning community has recently become aware of the poor replicability of many experimental ML studies. Among other reasons, the large computational cost of some of these experiments, notably in reinforcement learning and in large language models, makes most of these experiments non replicable. This cost prompts research towards comparison methods that require as few computations as possible to obtain a replicable conclusion. Aside the computational cost, the conclusion of the comparison should also be replicable which calls for appropriate statistical tests. AdaStop is a recently introduced statistical test based on multiple group sequential tests. AdaStop adapts the number of executions of each experiment to stop as early as possible while ensuring that enough information is available to distinguish algorithms that perform better than the others in a statistically significant way. AdaStop has been initially exemplified on reinforcement learning tasks. In this short paper, we give new use cases of AdaStop beyond reinforcement learning to provide the computer science community with a tool that can be used to compare algorithms in any domain requiring computational studies.

8.4.2 Algorithmic auditing

Active Fourier Auditor for Estimating Distributional Properties of ML Models, [29]

With the pervasive deployment of Machine Learning (ML) models in real-world applications, verifying and auditing properties of ML models have become a central concern. In this work, we focus on three properties: robustness, individual fairness, and group fairness. We discuss two approaches for auditing ML model properties: estimation with and without reconstruction of the target model under audit. Though the first approach is studied in the literature, the second approach remains unexplored. For this purpose, we develop a new framework that quantifies different properties in terms of the Fourier coefficients of the ML model under audit but does not parametrically reconstruct it. We propose the Active Fourier Auditor (AFA), which queries sample points according to the Fourier coefficients of the ML model, and further estimates the properties. We derive high probability error bounds on AFA's estimates, along with the worst-case lower bounds on the sample complexity to audit them. Numerically we demonstrate on multiple datasets and models that AFA is more accurate and sample-efficient to estimate the properties of interest than the baselines.

How Much Does Each Datapoint Leak Your Privacy? Quantifying the Per-datum Membership Leakage, [34]

We study the per-datum Membership Inference Attacks (MIAs), where an attacker aims to infer whether a fixed target datum has been included in the input dataset of an algorithm and thus, violates privacy. First, we define the membership leakage of a datum as the advantage of the optimal adversary targeting to identify it. Then, we quantify the per-datum membership leakage for the empirical mean, and show that it depends on the Mahalanobis distance between the target datum and the data-generating distribution. We further assess the effect of two privacy defences, i.e. adding Gaussian noise and sub-sampling. We quantify exactly how both of them decrease the per-datum membership leakage. Our analysis builds on a novel proof technique that combines an Edgeworth expansion of the likelihood ratio test and a Lindeberg-Feller central limit theorem. Our analysis connects the existing likelihood ratio and scalar product attacks, and also justifies different canary selection strategies used in the privacy auditing literature. Finally, our experiments demonstrate the impacts of the leakage score, the sub-sampling ratio and the noise scale on the per-datum membership leakage as indicated by the theory.

Testing Credibility of Public and Private Surveys through the Lens of Regression, [37]

Testing whether a sample survey is a credible representation of the population is an important question to ensure the validity of any downstream research. While this problem, in general, does not have

an efficient solution, one might take a task-based approach and aim to understand whether a certain data analysis tool, like linear regression, would yield similar answers both on the population and the sample survey. In this paper, we design an algorithm to test the credibility of a sample survey in terms of linear regression. In other words, we design an algorithm that can certify if a sample survey is good enough to guarantee the correctness of data analysis done using linear regression tools. Nowadays, one is naturally concerned about data privacy in surveys. Thus, we further test the credibility of surveys published in a differentially private manner. Specifically, we focus on Local Differential Privacy (LDP), which is a standard technique to ensure privacy in surveys where the survey participants might not trust the aggregator. We extend our algorithm to work even when the data analysis has been done using surveys with LDP. In the process, we also propose an algorithm that learns with high probability the guarantees a linear regression model on a survey published with LDP. Our algorithm also serves as a mechanism to learn linear regression models from data corrupted with noise coming from any subexponential distribution. We prove that it achieves the optimal estimation error bound for ℓ_1 linear regression, which might be of broader interest. We prove the theoretical correctness of our algorithms while trying to reduce the sample complexity for both public and private surveys. We also numerically demonstrate the performance of our algorithms on real and synthetic datasets.

8.4.3 ML for Science and Optimization

Bistability in the sunspot cycle, [17]

A direct dynamical test of the sunspot cycle is carried out to indicate that a stochastically forced nonlinear oscillator characterizes its dynamics. The sunspot series is then decomposed into its eigen time-delay coordinates. The relevant analysis reveals that the sunspot series exhibits bistability, with the possibility of modeling the solar cycle as a stochastically and periodically forced bistable oscillator, accounting for poloidal and toroidal modes of the solar magnetic field. Such a representation enables us to conjecture stochastic resonance as the key mechanism in amplifying the planetary influence on the Sun, and that extreme events, due to turbulent convection noise inside the Sun, dictate crucial phases of the sunspot cycle, such as the Maunder minimum.

The Steepest Slope toward a Quantum Few-body Solution, [16]

Quantum few-body systems are deceptively simple. Indeed, with the notable exception of a few special cases, their associated Schr´odinger equation cannot be solved analytically for more than two particles. One has to resort to approximation methods to tackle quantum few-body problems. In particular, variational methods have been proposed to ease numerical calculations and obtain precise solutions. One such method is the Stochastic Variational Method, which employs a stochastic search to determine the number and parameters of correlated Gaussian basis functions used to construct an ansatz of the wave function. Stochastic methods, however, face numerical and optimization challenges as the number of particles increases. We introduce a family of gradient variational methods that replace stochastic search with gradient optimization. We comparatively and empirically evaluate the performance of the baseline Stochastic Variational Method, several instances of the gradient variational method family, and some hybrid methods for selected few-body problems. We show that gradient and hybrid methods can be more efficient and effective than the Stochastic Variational Method. We discuss the role of singularities, oscillations, and gradient optimization strategies in the performance of the respective methods.

IDEQ: an improved diffusion model for the TSP, [36]

We investigate diffusion models to solve the Traveling Salesman Problem. Building on the recent DIFUSCO and T2TCO approaches, we propose IDEQ (constrained Inverse Diffusion and EQuivalence class-based retraining of diffusion models for combinatorial optimization). IDEQ improves the quality of the solutions by leveraging the constrained structure of the state space of the TSP. Another key component of IDEQ consists in replacing the last stages of DIFUSCO curriculum learning by considering a uniform distribution over the Hamiltonian tours whose orbits by the 2-opt operator converge to the optimal solution as the training objective. Our experiments show that IDEQ improves the state of the art for such neural network based techniques on synthetic instances. More importantly, our experiments show that IDEQ performs very well on the instances of the TSPlib, a reference benchmark in the TSP community: it matches the performance of the best heuristics, LKH3, being even able to obtain better solutions than LKH3 on 2 instances of the TSPlib defined on 1577 and 3795 cities. IDEQ obtains 0.3% optimality gap on

TSP instances made of 500 cities, and 0.5% on TSP instances with 1000 cities. This sets a new SOTA for neural based methods solving the TSP. Moreover, IDEQ exhibits a lower variance and better scales-up with the number of cities with regards to DIFUSCO and T2TCO.

8.4.4 Robustness in ML

When Witnesses Defend: A Witness Graph Topological Layer for Adversarial Graph Learning, [33]

Capitalizing on the intuitive premise that shape characteristics are more robust to perturbations, we bridge adversarial graph learning with the emerging tools from computational topology, namely, persistent homology representations of graphs. We introduce the concept of witness complex to adversarial analysis on graphs, which allows us to focus only on the salient shape characteristics of graphs, yielded by the subset of the most essential nodes (i.e., landmarks), with minimal loss of topological information on the whole graph. The remaining nodes are then used as witnesses, governing which higher-order graph substructures are incorporated into the learning process. Armed with the witness mechanism, we design Witness Graph Topological Layer (WGTL), which systematically integrates both local and global topological graph feature representations, the impact of which is, in turn, automatically controlled by the robust regularized topological loss. Given the attacker's budget, we derive the important stability guarantees of both local and global topology encodings and the associated robust topological loss. We illustrate the versatility and efficiency of WGTL by its integration with five GNNs and three existing non-topological defense mechanisms. Our extensive experiments across six datasets demonstrate that WGTL boosts the robustness of GNNs across a range of perturbations and against a range of adversarial attacks, leading to relative gains of up to 18%.

8.4.5 Federated learning

Don't Forget What I did?: Assessing Client Contributions in Federated Learning, [39]

Federated Learning (FL) is a collaborative machine learning (ML) approach, where multiple clients participate in training an ML model without exposing the private data. Fair and accurate assessment of client contributions is an important problem in FL to facilitate incentive allocation and encouraging diverse clients to participate in a unified model training. Existing methods for assessing client contribution adopts co-operative game-theoretic concepts, such as Shapley values, but under simplified assumptions. In this paper, we propose a history-aware game-theoretic framework, called FLContrib, to assess client contributions when a subset of (potentially non-i.i.d.) clients participate in each epoch of FL training. By exploiting the FL training process and linearity of Shapley value, we develop FLContrib that yields a historical timeline of client contributions as FL training progresses over epochs. Additionally, to assess client contribution under limited computational budget, we propose a scheduling procedure that considers a two-sided fairness criteria to perform expensive Shapley value computation only in a subset of training epochs. In experiments, we demonstrate a controlled trade-off between the correctness and efficiency of client contributions assessed via FLContrib. To demonstrate the benefits of history-aware client contributions, we apply FLContrib to detect dishonest clients conducting data poisoning in FL training.

9 Bilateral contracts and grants with industry

Participants: Odalric-Ambrym Maillard, Philippe Preux, Debabrota Basu.

9.1 Bilateral contracts with industry

- contract with Ubisoft, 2023–2026: PI: O-A. Maillard.

  This contract is related to A. Kobanda's Ph.D. "Continual Reinforcement Learning with changing environments: Application to Video Games"

- contract with Lilly Group, 2023–2026, PI: Ph. Preux.

  This contract is related to M. Basson's Ph.D. "Reinforcement learning to solve combinatorial optimization problems".

- contract with Saint-Gobain Research, 2023–2026, PI: Ph. Preux.

  This contract is related to M. Basson's Ph.D. "Reinforcement learning for advanced control of industrial processus".

- contract with Bits2beat Analytics, 2024, PI: D. Basu.

  Collaboration contract on classification, segmentation and semantic interpretation of ECGs.

10 Partnerships and cooperations

Participants: Philippe Preux, Odalric-Ambrym Maillard, Emilie Kaufmann, Debabrota Basu, Rémy Degenne, Riadh Akrour.

10.1 International initiatives

10.1.1 Inria associate team not involved in an IIL or an international program

DC4SCM

Title: Data Collection for Smart Crop Management

Duration: 2020 → 2024

Coordinator: Chandan Panda (dr.ckpanda@gmail.com)

Partners:

- Bihar Agricultural University, Sabour, India (Inde)

Inria contact: Philippe Preux

Summary: Project DC4SCM focusses on data related to crop management of small farm holders, in developping countries. The goal is to investigate which data to collect, collect them, analyse them, and use them (beyond DC4SCM) to recommend practices. DC4SCM gathers 3 research teams: Bihar agricutural university in India bring its network of small farmholders et its expertise in agriculture; Inria Fun brings its expertise in the design, deployment, and management of sensor networks; Inria SequeL brings its expertise in machine learning and data science. For Bihar, DC4SCM provides the opportunity to gain skills in data science and IoT; for Fun, DC4SCM provides an opportunity to investigate a new type of applications of IoT; SequeL, is yet collaborating with CIRAD and CIAT on the use and development of machine learning to provide recommendations to farmers in Malawi: DC4SCM is the opportunity to obtain new and complementary data for this project. More generally, DC4SCM aims at contributing to the development of sustainable agriclture practices, in particular in developping countries.

RELIANT

Title: Real-life bandits

Duration: 2022 → 2024

Coordinator: Junya Honda (honda@i.kyoto-u.ac.jp)

Partners:

- Kyoto University Kyoto (Japon)

Inria contact: Odalric-Ambrym Maillard

Summary: The RELIANT project is about studying applicability to the real-world of sequential decision making from a reinforcement learning (RL) and multi-armed bandit (MAB) theory standpoint. Building on over a decade of leading expertise in advancing the field of MAB and RL theory, our two teams have also developed interactions with practitioners (e.g. in healthcare, personalized medicine or agriculture) in recent projects, in the quest to bring modern bandit theory to societal applications, for real. This quest for real-world reinforcement learning, rather than working in simulated and toyish environments is actually today's main grand-challenge of the field that hinders applications to the society and industry. While MABs are acknowledged to be the most applicable building block of RL, as experts interacting with practitioners from different fields we have identify a number of key bottlenecks on which joining our efforts is expected to significantly impact the applicability of MAB to the real-world. Those as related to the typically small samples size that arise in medical applications, the complicated type of rewards distributions that arise, e.g. in agriculture, the numerous constraints (such as fairness) that should be taken into account to speed up learning and make ethical decisions, and the possible non-stationary aspects of the tasks. We suggest to connect on the mathematical level our complementary expertise on multi-armed bandit (MAB), sequential hypothesis testing (SHT) and Markov decision processes (MDP) to address these challenges and significantly advance the design of the next generation of sequential decision making algorithms for real-life applications.

10.2 International research visitors

10.2.1 Visits of international scientists

Other international visits to the team

Riccardo Poiani

Status PhD

Institution of origin: Politecnico Milano

Country: Italy

Dates: March 4th-May 31st

Context of the visit: Research visit during his PhD

Mobility program/type of mobility: research stay partially founded by the MOBLILEX program

10.2.2 Visits to international teams

Research stays abroad

Marc Jourdan

Visited institution: Università degli Studi di Milano

Country: Italy

Dates: April 1st- June 30th

Context of the visit: research visit during his PhD

Mobility program/type of mobility: research stay partially founded by the IKS mobility grant

Achraf Azize

Visited institution: University Kyoto

Country: Japan

Dates: September 13th - November 13th

Context of the visit: research visit during his PhD and within the RELIANT associate team.

Mobility program/type of mobility: research stay funded by RELIANT associate team.

Debabrota Basu

Visited institution: Indian Statistical Institute, Kolkata

Country: India

Dates: June-July

Context of the visit: research visit to co-supervise students and leading to an associate team application under Inria-DST join project calls.

Mobility program/type of mobility: research stay funded by Indian Statistical Institute.

Debabrota Basu

Visited institution: Bihar Agricultural University, Sabour

Country: India

Dates: June

Context of the visit: research visit to progress in the DC4SCM project and finalising publishable results from the project.

Mobility program/type of mobility: visit funded by DC4SCM associate team.

10.3 European initiatives

10.3.1 Other european programs/initiatives

Title: CausalXRL

Partner Institutions:

- University of Sheffield, Department of Computer Science
- University of Vienna, Faculty of Computer Science
- Inria, Scool, Lille, France

Date/Duration: 48 months

Additionnal info/keywords: Causality, Reinforcement Learning, Explanations.

10.4 National initiatives

10.4.1 ANR projects

Scool is involved in 4 ANR projects:

- ANR JCJC FATE, PI: R. Degenne, 2023–2027
- ANR JCJC REPUBLIC, PI: D. Basu, 2023–2026
- ANR BIP-UP, partnership: Scool/Inserm (CHU de Lille), PI: Ph. Preux, 2023–2026.
- ANR JCJC NeuRL, PI: R. Akrour, 2024–2028

10.4.2 PEPR projects

Scool is involved in 2 PEPR:

- PEPR AI: project FOUNDRY, local head: E. Kaufmann (description below);
- PEPR « Agroécologie et numérique », Pl@ntAgroEco, local head: O.-A. Maillard.

Title: FOUNDRY

Duration: July 2024 → June 2028

Coordinator: Panayotis Mertikopoulous, Polaris, Univ. Grenoble Alpes

Partners:

- POLARIS: a joint research team between the CNRS, Inria, and Univ. Grenoble Alpes.
- ENS Lyon: faculty from the pure and applied mathematics department of ENS Lyon.
- Inria FAIRPLAY: a joint team between Criteo, IP Paris (ENSAE and Ecole Polytechnique), and Inria.
- LTCI: the informations and communications laboratory of Télécom Paris.
- MILES: the machine intellligence and learning systems of the LAMSADE lab at Paris Dauphine.
- Inria Scool

Inria contact: Emilie Kaufmann

Summary: From automated hospital admission systems powered by machine learning (ML), to flexible chatbots capable of fluent conversations and self-driving cars, the wildfire spread of artificial intelligence (AI) has brought to the forefront a crucial question with far-reaching ramifications for the society at large: Can ML systems and models be relied upon to provide trustworthy output in high-stakes, mission- critical environments? These questions invariably revolve around the notion of *robustness*, an operational desideratum that has eluded the field since its nascent stages. One of the main reasons for this is the fact that ML models and systems are typically data-hungry and highly sensitive to their training input, so they tend to be brittle, narrow-scoped, and unable to adapt to situations that go beyond their training envelope. On that account, the core vision of the proposed research is that robustness cannot be achieved by blindly throwing more data and computing power to larger and larger models with exponentially growing energy requirements (and a commensurate carbon footprint to boot). Instead, our proposal intends to rethink and develop the core theoretical and methodological FOUNDations of Robustness and reliabilitY (FOUNDRY) that are needed to build and instill trust in ML-powered technologies and systems from the ground up.

Title: Pl@ntAgroEco

Duration: July 2024 → June 2028

Coordinator: Alexis Joly, Inria Zenith, and Pierre Bonnet CIRAD, AMAP.

Partners:

- INRAE
- INRIA
- IRD
- CIRAD
- Tela Botanica
- Université de Montpellier

- Université Paris-Saclay

Inria contact: Odalric-Ambrym Maillard

Summary: Agroecology necessarily involves crop diversification, but also the early detection of diseases, deficiencies and stresses (hydric, etc.), as well as better management of biodiversity. The main stumbling block is that this paradigm shift in agricultural practices requires expert skills in botany, plant pathology and ecology that are not generally available to those working in the field, such as farmers or agri-food technicians. Digital technologies, and artificial intelligence in particular, can play a crucial role in overcoming this barrier to access to knowledge.

The aim of the Pl@ntAgroEco project will be to design, experiment with and develop new high-impact agro-ecology services within the Pl@ntNet platform. This includes :

- research in AI and plant sciences ;
- agile development of new components within the platform;
- organization of participatory science programs and animation of the Pl@ntNet user community.

Ce programme de travail a pour but de produire une amélioration de la détection et reconnaissance des maladies végétales, de l'identification des niveaux infraspécifiques. Il permettra le développement d'outils d'estimation de la sévérité des symptômes, carences, stades de déclin et stress hydrique ou de caractérisation des associations d'espèces à partir d'images multi-spécimens. Il améliorera la connaissance des espèces.

Le projet Pl@ntAgroEco rassemble des forces complémentaires en matière de recherche, de développement et d'animation. S'ajouteront à l'équipe pluridisciplinaire chargée de la plateforme Pl@ntNet de nouvelles forces de recherche ayant une expertise reconnue dans les sciences participatives. Le consortium rassemblera 10 partenaires incluant des organismes de recherche, des universités, des acteurs de la société civile et des partenaires internationaux.

10.4.3 Other projects in France

Scool is involved in the Regalia pilot-project.

Other collaborations:

- L. Richert, R. Thiébaut, Inria SISTM, Bordeaux, bandits for vaccine clinical trials.
- W. M. Koolen, CWI Amsterdan & University of Twente, concentration of information divergences.

10.5 Regional initiatives

- O.-A. Maillard and Ph. Preux are supported by an AI chair. 3/5 of this chair is funded by the Metropole Européenne de Lille, the other 2/5 by the Université de Lille and Inria, through the AI Ph.D. ANR program. 2020–2024.

  This chair is dedicated to the advancement of research on reinforcement learning.

- Collaboration with U. INSERM 1190/Université de Lille/CHU de Lille with Prof. F. Pattou, ANR BIP-UP (É. Kaufmann, T. Mathieu, O-A. Maillard, Ph. Preux).
- Collaboration with Prof. E. Heymann (Université de Lille, UREPSS) on the prediction of the influence of physical activities on diabetes in teenagers (Ph. Preux).
- Collaboration with J-B. Colliat at the LamCube, mechanical engineering UMR at the Université de Lille on the use of reinforcement learning for smart testing (Ph. Preux).
- D. Basu and Ph. Preux are participants in the Decision-making Processes under Extreme Radical Uncertainties (DePERU) project funded by an AAP Cross Disciplinary grant. It is a collaboration between economists, psychologists, social scientists, and computer scientists in Lille region to understand human decision making under extreme uncertainties.

11 Dissemination

Participants: Juliette Achddou, Riadh Akrour, Debabrota Basu, Rémy Degenne, Emilie Kaufmann, Hector Kohler, Odalric-Ambrym Maillard, Timothée Mathieu, Philippe Preux.

11.1 Promoting scientific activities

11.1.1 Scientific events: organisation

- H. Kohler co-organized the Workshop on Interpretable Policies in Reinforcement Learning (InterPol) at the Reinforcement Learning Conference, Amherst, USA, August 24.

Member of the organizing committees

- O.-A. Maillard: Session organizer at RL4SN Informs conference "Challenges and progresses in Statistical Reinforcement Learning", June 17-21 2024, Toulouse.

11.1.2 Scientific events: selection

Member of the conference program committees

- R. Degenne: member of the Senior PC of the conference Algorithmic Learning Theory (ALT).
- E. Kaufmann: member of the Senior PC of the conference Algorithmic Learning Theory (ALT).
- Ph. Preux: PC IJCAI, PC ECML.
- O.-A. Maillard: member of the Senior PC of the Conference On Learning Theory (COLT).
- D. Basu: PC of AAAI, AAMAS, PRICAI.

Reviewer

- J. Achddou : reviewer for ALT, AISTATS.
- R. Akrour : reviewer for NeurIPS.
- R. Degenne: reviewer for COLT, NeurIPS, ALT.
- E. Kaufmann: reviewer for ICML, ALT.
- T. Mathieu: reviewer for COLT, AISTATS.
- D. Basu: reviewer of NeuRIPS, ICML, AISTATS, IJCAI, ICLR, UAI.

11.1.3 Journal

Member of the editorial boards

- O.-A. Maillard: editor for the Journal of Machine Learning Research.

Reviewer - reviewing activities

- R. Degenne: reviewer for the Journal of Machine Learning Research.
- E. Kaufmann: reviewer for the Journal of Machine Learning Research.
- T. Mathieu: reviewer for the Journal of the Royal Statistical Society: Series B, Annales de l'Institut Henri Poincaré, Journal of the American Statistical Association.
- O.-A. Maillard: reviewer for the Journal of Machine Learning Research, for Mathematics of Operation Research, for Statistics and Probability Letters.
- D. Basu: reviewer of JMLR, TMLR, Entropy, IEEE Transactions on Artificial Intelligence (TAI), Transactions on Signal Processing (TSP), IEEE Transactions on Parallel and Distributed Systems (TPDS), IEEE Trans actions on Automatic Control (TAC), IEEE Access, IEEE Transactions on Information Forensics & Security (TIFS), IEEE Transactions on Dependable & Secure Computing (TDSC).

11.1.4 Invited talks

- R. Akrour: talk on preference-based reinforcement learning at Journées Synthèse de Programmes in Bordeaux.
- T. Mathieu: talk on robust statistics in Séminaire de Statistiques de Montpellier, talk on robust statistics in Séminaire parisien de statistique SemStat.
- O.-A. Maillard: Invited speaker at Inria-Brasil workshop on Digital Agriculture, September 10-11 2024, Montpellier (France), at W'Happy Digital Day "IA et Agroécologie", May 16 2024, Tournai (Belgium), at Table Ronde DigitAgora "Le numérique en agriculture", April 24 2024, Montpellier (France).
- D. Basu: Invited talks: (i) When Privacy Meets Partial Information: Privacy-Utility Trade-offs in Reinforcement Learning, IndoML, India (keynote). (ii) Auditing Bias of ML Systems: An Algorithmic Montage, Digital regulation: the contribution of science, Bruxelles, EU (keynote). (iii) Challenges of Data-centrism, Workshop on Data-Driven Futures in the Age of AI: Leveraging Insights and Addressing Challenges, ADRI, India (keynote). (iv) Panel Discussion: A Multidisciplinary View of Responsible Data-Centrism, ADRI, India (Public Talk). (v) The fair game: auditing and debiasing AI algorithms over time, JUST-AI Summer Colloquium on Individual Safeguards in the Era of AI by EU & University of Liége. (vi) Sequential Decision Making under Partial Information: RL, Uncertainty Quantification, & Responsible AI, Séminaire doctoral – ED SHS Université de Lille. (vii) The Privacy Game: Attacks with and Defenses for Online ML Al- gorithms, Workshop Inria-University of Waterloo–Université de Bordeaux, Bordeaux, France. (viii) AI for Agricultural Sciences, CAFT training program on "Harnessing Disruptive Technology in ICT for Agricultural Extension and Research", Sabour, India (Public talk).

11.1.5 Leadership within the scientific community

- D. Basu: selected as a fellow of European Laboratory of Learning and Intelligent Systems (ELLIS) Society.

11.1.6 Scientific expertise

- Ph. Preux: evaluation of an ANR project.
- D. Basu: evaluation of an ANR grant application.

11.1.7 Research administration

- Ph. Preux:
  - scientific coordinator of CPER CornelIA
  - member of the CSS5 (data science and models) at IRD
  - member of the scientific committee of the MathNum department at Inrae
  - member of the scientific committee of PEPR agroécologie et numérique
  - member of the scientific and ethical committee of INCLUDE (data warehouse of CHU Lille)
  - member of the DAS[1] health at Région Hauts-de-France
  - member of the BCEP at Inria Lille.

11.2 Teaching - Supervision - Juries

11.2.1 Teaching

- J. Achddou: « Algorithmique Numérique pour l'Optimisation », M2 in Computer Science and Statistics, Polytech Lille.
- J. Achddou: « Algorithmique Numérique pour l'Optimisation », M1 in Computer Science and Statistics, Polytech Lille.
- R. Akrour: Sequential Decision Making, M2 in Data Science, Centrale Lille and Université de Lille
- R. Akrour: « Option découverte: Machine Learning », L3 in Computer Science, Université de Lille
- R. Akrour: « Perception et motricité 2 », L2 MIASHS, Université de Lille
- R. Degenne: « Sequential learning », M2 MVA, ENS Paris-Saclay
- R. Degenne: « Sequential learning », Centrale Lille
- D. Basu: « Sequential Decision Making », M2 in Data Science, Centrale Lille and Université de Lille
- D. Basu: « Research Reading Group », M2 in Data Science, Centrale Lille and Université de Lille
- D. Basu: « Advanced Machine Learning and Decision Making », Centrale Lille
- E. Kaufmann: « Statistics 2 », M1 Data Science, Ecole Centrale Lille.
- Ph. Preux: « Prise de décision séquentielle dans l'incertain », M2 in Computer Science, Université de Lille.
- Ph. Preux: « Apprentissage par renforcement », M2 in Computer Science, Université de Lille.
- Ph. Preux: « Science des données II », L3 MIASHS, Université de Lille.
- Ph. Preux: « Réseaux de neurones », L1 Maths-Informatique, Université de Lille.
- Ph. Preux: « IA et apprentissage automatique », DU IA & Santé, Université de Lille.
- O-A. Maillard: « Reinforcement Learning Research Challenges », Executive Master Ecole Polytechnique.

[1] strategic activity domain

11.2.2 Supervision

- Ph. Preux:
  - C. Blondelle, L3 MIASHS
  - Q. Uguen, Centrale-Lille
  - Ph.D. students: M. Centa, M. Basson, P. Saux,
- Ph. Preux and R. Akrour: Ph.D. students: H. Kohler, Y. Berthelot
- Ph. Preux and O-A. Maillard: Ph.D. student: P. Saux
- Ph. Preux and D. Basu: Ph.D. students: A. Achraf, A. Ajarra
- E. Kaufmann and R. Degenne: PhD students: M. Jourdan, A. Tuynman
- E. Kaufmann and D. Basu: PhD student: T. Michel
- E. Kaufmann: PhD student: C. Kone (with Laura Richert, Université de Bordeaux, Inria SISTM)
- R. Akrour: PhD student: Brahim Driss.
- R. Akrour: Research engineer: A. Davey.
- R. Akrour: Masters student: M. Kabouri (with Joni Pajarinen, Aalto University, Finland).
- O-A. Maillard: Ph.D. students: S. Vashishtha, A. Kobanda.
- O-A. Maillard and D. Basu: Ph.D. student: U. Das.
- O-A. Maillard: Research engineer: W. Radji (until Sep 2024), then as Ph.D student.
- O-A. Maillard: Research engineer: H. Carvajal
- O-A. Maillard and T. Mathieu: Master student: A. Prevost (until Sep 2024), then as Ph.D. student.
- O-A. Maillard: Postdoc: T. Dan (until Sep 2024).
- O-A. Maillard: Postdoc: T. Lefort.
- D. Basu: Masters student: A. Olivier (with Bits2beat, Bordeaux).
- D. Basu: PhD student: E. Jorge (with C. Dimitrakakis, Chalmers Univ. of Technology, Sweden).
- D. Basu: Research engineer: G. Pourcel.
- R. Degenne: Master student: L. Luccioli.

11.2.3 Juries

- Ph. Preux:
  - CR and DR IRD CSS5
  - Ph.D. defenses: P. Saux (Lille, co-supervisor), É. Arnaud (Amiens, reviewer), J. Bujalance Martin (PSL/Mines, reviewer), A. Azize, (Lille, co-supervisor), I. Harith (IMT, chair), M. Zadem (ENSTA, reviewer)
- E. Kaufmann:
  - Ph.D. defenses: A. Pacaud (Telecom Paris), P. Wang (KTH, Stockholm), A. Chaouki (Ecole Polytechnique), M. Jourdan, (Lille, co-supervisor), I. Harith (IMT, chair), J. Whitehouse (CMU, Pittsburgh)
- O-A. Maillard:

  - Ph.D. defenses: A. Barrier (ENS Lyon, reviewer), D. Brellmann (Telecom Paris, reviewer), T. Lefort (Univ. Montpellier), D. Delande, (Univ. Toulouse).
  - HdR defense: A. Cleynen (Univ. Montpellier, reviewer).
  - CSI: F. Fabre (Univ. Réunion), T. Delliaux (Univ. Toulouse).
- D. Basu:
  - Ph.D. defense: A. Azize (Lille, co-supervisor)
- R. Degenne:
  - Ph.D. defense: M. Jourdan (Lille, co-supervisor)

11.3 Popularization

11.3.1 Others science outreach relevant activities

- Ph. Preux:
  - 1/2/2024 : « l'intelligence artificielle, qu'en est-il ? », université populaire de Lille, Palais des Beaux-Arts, Lille.
  - 29/11/2024 : « L'intelligence artificielle. De quoi parle-t-on ? », public library Saint-Omer.
- J. Teigny:
  - « Introduction à l'alignement des IA »: at Inria Lille, journée du réseau métier Min2rien, CRIStAL, Science Po Lille.
  - « Introduction a l'intelligence artificielle », association Tourcoing Loisir Sortir
  - 4 sessions of co-animation of a game related to AI for high school students.

12 Scientific production

12.1 Major publications

[1] B. Balle and O.-A. Maillard. 'Spectral Learning from a Single Trajectory under Finite-State Policies'. In: *International conference on Machine Learning*. Proceedings of the International conference on Machine Learning. Sidney, France, July 2017. URL: https://hal.archives-ouvertes.fr/hal-01590940.

[2] D. Baudry, R. Gautron, E. Kaufmann and O.-A. Maillard. 'Optimal Thompson Sampling strategies for support-aware CVaR bandits'. In: 38th International Conference on Machine Learning. proceedings of machine learning research. Virtual, United States, 18th July 2021. URL: https://hal.science/hal-03447244.

[3] G. Dulac-Arnold, L. Denoyer, P. Preux and P. Gallinari. 'Sequential approaches for learning datum-wise sparse representations'. In: *Machine Learning* 89.1-2 (1st Oct. 2012), pp. 87–122. DOI: 10.1007/s10994-012-5306-7. URL: https://hal.inria.fr/hal-00747724.

[4] Y. Flet-Berliac and P. Preux. 'Only Relevant Information Matters: Filtering Out Noisy Samples to Boost RL'. In: *IJCAI 2020 - International Joint Conference on Artificial Intelligence*. Yokohama, Japan, July 2020. DOI: 10.24963/ijcai.2020/376. URL: https://hal.inria.fr/hal-02091547.

[5] B. Ghosh, D. Basu and K. S. Meel. 'Justicia: A Stochastic SAT Approach to Formally Verify Fairness'. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI Conference on Artificial Intelligence. Vol. 35. Proceedings of the AAAI Conference on Artificial Intelligence 9. Virtual, Canada, Feb. 2021, pp. 7554–7563. URL: https://hal.science/hal-03445831.

[6] M. Jourdan, R. Degenne, D. Baudry, R. de Heide and E. Kaufmann. 'Top Two Algorithms Revisited'. In: NeurIPS 2022 - 36th Conference on Neural Information Processing System. Advances in Neural Information Processing Systems. New Orleans, United States, 28th Nov. 2022. URL: https://hal.science/hal-03825103.

[7] H. Kadri, E. Duflos, P. Preux, S. Canu, A. Rakotomamonjy and J. Audiffren. 'Operator-valued Kernels for Learning from Functional Response Data'. In: *Journal of Machine Learning Research* 17.20 (2016), pp. 1–54. URL: https://hal.archives-ouvertes.fr/hal-01221329.

[8] E. Kaufmann, P. Ménard, O. Darwiche Domingues, A. Jonsson, E. Leurent and M. Valko. 'Adaptive reward-free exploration'. In: Algorithmic Learning Theory. Paris, France, 2021. URL: https://hal.science/hal-02864574.

[9] O.-A. Maillard. 'Boundary Crossing Probabilities for General Exponential Families'. In: *Mathematical Methods of Statistics* 27 (2018). URL: https://hal.archives-ouvertes.fr/hal-01737150.

[10] O.-A. Maillard, H. Bourel and M. S. Talebi. 'Tightening Exploration in Upper Confidence Reinforcement Learning'. In: *International Conference on Machine Learning*. Vienna, Austria, July 2020. URL: https://hal.archives-ouvertes.fr/hal-03000664.

[11] O. Nicol, J. Mary and P. Preux. 'Improving offline evaluation of contextual bandit algorithms via bootstrapping techniques'. In: *International Conference on Machine Learning*. Ed. by E. Xing and T. Jebara. Vol. 32. Journal of Machine Learning Research, Workshop and Conference Proceedings; Proceedings of The 31st International Conference on Machine Learning. Beijing, China, June 2014. URL: https://hal.inria.fr/hal-00990840.

[12] F. Pesquerel and O.-A. Maillard. 'IMED-RL: Regret optimal learning of ergodic Markov decision processes'. In: NeurIPS 2022 - Thirty-sixth Conference on Neural Information Processing Systems. Thirty-sixth Conference on Neural Information Processing Systems. New-Orleans, United States, 28th Nov. 2022. URL: https://hal.science/hal-03825423.

[13] F. Strub, M. Seurin, E. Perez, H. De Vries, J. Mary, P. Preux, A. Courville and O. Pietquin. 'Visual Reasoning with Multi-hop Feature Modulation'. In: *ECCV 2018 - 15th European Conference on Computer Vision*. Ed. by V. Ferrari, M. Hebert, C. Sminchisescu and Y. Weiss. Vol. 11205-11220. Part of the Lecture Notes in Computer Science book series - LNCS 11209. Munich, Germany, Sept. 2018, pp. 808–831. URL: https://hal.archives-ouvertes.fr/hal-01927811.

12.2 Publications of the year

International journals

[14] T. Mathieu, D. Basu and O.-A. Maillard. 'Bandits with Stochastic Corruption: Lower Bounds on Regret and Robust Optimistic Algorithms'. In: *Transactions on Machine Learning Research Journal* (Jan. 2024). URL: https://hal.science/hal-04615733 (cit. on p. 10).

[15] T. Mathieu, R. Della Vecchia, A. Shilova, M. Medeiros Centa, H. Kohler, O.-A. Maillard and P. Preux. 'AdaStop: adaptive statistical testing for sound comparisons of Deep RL agents'. In: *Transactions on Machine Learning Research Journal* (2024). URL: https://inria.hal.science/hal-04132861 (cit. on p. 12).

[16] P. Recchia, D. Basu, M. Gattobigio, C. Miniatura and S. Bressan. 'The Steepest Slope toward a Quantum Few-body Solution: Gradient Variational Methods for the Quantum Few-body Problem'. In: *Few-Body Systems* 65.4 (19th Nov. 2024), p. 102. DOI: 10.1007/s00601-024-01965-7. URL: https://hal.science/hal-04672894 (cit. on p. 15).

[17] S. Vashishtha and K. Sreenivasan. 'Bistability in the sunspot cycle'. In: *EPL - Europhysics Letters* 148.2 (14th Oct. 2024), p. 23001. DOI: 10.1209/0295-5075/ad7f85. URL: https://hal.science/hal-04809100. In press (cit. on p. 15).

International peer-reviewed conferences

[18] S. Agrawal, T. Mathieu, D. Basu and O.-A. Maillard. 'CRIMED: Lower and Upper Bounds on Regret for Bandits with Unbounded Stochastic Corruption'. In: *Proceedings of Machine Learning Research*. International Conference on Algorithmic Learning Theory (ALT). Vol. 237. San Diego (CA), United States, Mar. 2024, pp. 74–124. URL: https://hal.science/hal-04260464 (cit. on p. 10).

[19] A. Azize and D. Basu. 'Concentrated Differential Privacy for Bandits'. In: 2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML). Toronto, Canada: IEEE, 9th Apr. 2024, pp. 78–109. DOI: 10.1109/SaTML59370.2024.00013. URL: https://hal.science/hal-04611650 (cit. on p. 11).

[20] A. Azize and D. Basu. 'Open Problem: What is the Complexity of Joint Differential Privacy in Linear Contextual Bandits?' In: Proceedings of Thirty Seventh Conference on Learning Theory. PMLR. Edmonton (Alberta), Canada, July 2024. URL: https://hal.science/hal-04621903 (cit. on p. 11).

[21] E. Carlsson, D. Basu, F. D. Johansson and D. Dubhashi. 'Pure Exploration in Bandits with Linear Constraints'. In: *Proceedings of Machine Learning Research (PMLR)*. International Conference on Artificial Intelligence and Statistics. Vol. 238. Proceedings of Machine Learning Research (PMLR). Valencia (Espagne), Spain, May 2024, pp. 334–342. URL: https://hal.science/hal-04203235 (cit. on p. 9).

[22] T. Q. T. Dam, O.-A. Maillard and E. Kaufmann. 'Power Mean Estimation in Stochastic Monte-Carlo Tree Search'. In: *40th conference on Uncertainty in Artificial Intelligence*. Uncertainty in Artificial Intelligence. Barcelona, Spain, 5th July 2024. URL: https://inria.hal.science/hal-04714124 (cit. on p. 9).

[23] M. Kallel, D. Basu, R. Akrour and C. d'Eramo. 'Augmented Bayesian Policy Search'. In: The Twelfth International Conference on Learning Representations (ICLR). Vienna, Austria, May 2024. URL: https://hal.science/hal-04616536 (cit. on p. 13).

[24] H. Kohler, Q. Delfosse, R. Akrour, K. Kersting and P. Preux. 'Interpretable and Editable Programmatic Tree Policies for Reinforcement Learning'. In: European Workshop on Reinforcement Learning. Vol. 17. Toulouse, France, 2024. URL: https://inria.hal.science/hal-04784919 (cit. on p. 12).

[25] R. Poiani, R. Degenne, E. Kaufmann, A. M. Metelli and M. Restelli. 'Optimal Multi-Fidelity Best-Arm Identification'. In: Advances in Neural Information Processing Systems (NeurIPS). Vancouver (BC), Canada, 10th Dec. 2024. URL: https://hal.science/hal-04811199 (cit. on p. 12).

[26] H. Saber and O.-A. Maillard. 'Bandits with Multimodal Structure'. In: *Reinforcement Learning Conference (Reinforcement Learning Journal)*. RLC 2024 - Reinforcement Learning Conference. Vol. 1. 5. Amherst Massachusetts, United States, 2024, p. 39. URL: https://inria.hal.science/hal-04711994 (cit. on p. 9).

[27] A. Shukla and D. Basu. 'Preference-based Pure Exploration'. In: Advances in Neural Information Processing Systems (NeurIPS). Vancouver (CA), Canada, Dec. 2024. URL: https://hal.science/hal-04733134 (cit. on p. 9).

[28] A. Tuynman, R. Degenne and E. Kaufmann. 'Finding good policies in average-reward Markov Decision Processes without prior knowledge'. In: NeurIPS. Vancouver (Canada), Canada, 10th Dec. 2024. URL: https://inria.hal.science/hal-04809429 (cit. on p. 8).

Conferences without proceedings

[29] A. Ajarra, B. Ghosh and D. Basu. 'Active Fourier Auditor for Estimating Distributional Properties of ML Models'. In: NeurIPS 2024 Workshop on Regulatable ML. Vancouver (BC), Canada, Dec. 2024. URL: https://hal.science/hal-04733059 (cit. on p. 14).

[30] U. Das and D. Basu. 'Learning to Explore with Lagrangians for Bandits under Unknown Linear Constraints'. In: Seventeenth European Workshop on Reinforcement Learning (EWRL). Toulouse, France, 24th Oct. 2024. URL: https://hal.science/hal-04784911 (cit. on p. 10).

[31] H. Eriksson, T. Tram, D. Basu, M. Alibeigi and C. Dimitrakakis. 'Reinforcement Learning in the Wild with Maximum Likelihood-based Model Transfer'. In: 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS). Auckland, New Zealand: ACM, May 2024, pp. 516–524. DOI: 10.5555/3635637.3662902. URL: https://hal.science/hal-04260795 (cit. on p. 13).

Edition (books, proceedings, special issue of a journal)

[32] *Statistical comparison in empirical computer science with minimal computation usage*. ACM REP '24: ACM Conference on Reproducibility and Replicability. Vol. 35. ACM, 18th June 2024, pp. 20–24. DOI: 10.1145/3641525.3663618. URL: https://inria.hal.science/hal-04718314 (cit. on p. 14).

Reports & preprints

[33] N. A. Arafat, D. Basu, Y. Gel and Y. Chen. *When Witnesses Defend: A Witness Graph Topological Layer for Adversarial Graph Learning*. 21st Sept. 2024. URL: https://inria.hal.science/hal-04708183 (cit. on p. 16).

[34] A. Azize and D. Basu. *How Much Does Each Datapoint Leak Your Privacy? Quantifying the Per-datum Membership Leakage*. 15th Feb. 2024. URL: https://hal.science/hal-04615701 (cit. on p. 14).

[35] A. Azize, M. Jourdan, A. A. Marjani and D. Basu. *Differentially Private Best-Arm Identification*. 10th June 2024. URL: https://hal.science/hal-04615690 (cit. on p. 11).

[36] M. Basson and P. Preux. *IDEQ: an improved diffusion model for the TSP*. RR-9558. INRIA Lille - Nord Europe, July 2024. URL: https://hal.science/hal-04778946 (cit. on p. 15).

[37] D. Basu, S. Chakraborty, D. Chanda, B. D. Das, A. Ghosh and A. Ray. *Testing Credibility of Public and Private Surveys through the Lens of Regression*. 7th Oct. 2024. URL: https://hal.science/hal-04733076 (cit. on p. 14).

[38] S. Chakraborty, S. Roy and D. Basu. *FLIPHAT: Joint Differential Privacy for High Dimensional Sparse Linear Bandits*. 22nd May 2024. URL: https://hal.science/hal-04615697 (cit. on p. 11).

[39] B. Ghosh, D. Basu, F. Huazhu, W. Yuan, R. Kanagavelu, J. J. Peng, L. Yong, G. S. M. Rick and W. Qingsong. *Don't Forget What I did?: Assessing Client Contributions in Federated Learning*. 11th Mar. 2024. URL: https://hal.science/hal-04615711 (cit. on p. 16).

[40] C. Han, D. Basu, M. Mangan, E. Vasilaki and A. Gilra. *Dynamical-VAE-based Hindsight to Learn the Causal Dynamics of Factored-POMDPs*. 12th Nov. 2024. URL: https://hal.science/hal-04785076 (cit. on p. 13).

[41] R. M. Nkhumise, D. Basu, T. J. Prescott and A. Gilra. *Measuring Exploration in Reinforcement Learning via Optimal Transport in Policy Space*. 14th Feb. 2024. URL: https://hal.science/hal-04702986 (cit. on p. 13).

[42] A. Shilova, T. Delliaux, P. Preux and B. Raffin. *Learning HJB Viscosity Solutions with PINNs for Continuous-Time Reinforcement Learning*. RR-9541. Inria Lille - Nord Europe, CRIStAL - Centre de Recherche en Informatique, Signal et Automatique de Lille - UMR 9189; Univ. Lille, CNRS, Centrale Lille, Inria UMR 9189 - CRIStAL,INRIA Lille Nord Europe, Villeneuve d'Ascq, France; Univ. Grenoble Alps, CNRS, Inria, Grenoble INP, LIG, 38000 Grenoble, France, 7th Feb. 2024, pp. 1–30. URL: https://inria.hal.science/hal-04445160 (cit. on p. 12).

12.3 Cited publications

[43] M. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, 1994 (cit. on p. 4).

[44] B. Recht. 'A Tour of Reinforcement Learning: The View from Continuous Control'. arxiv preprint 1806.09460. 2018 (cit. on p. 4).

[45] R. Sutton and A. Barto. *Reinforcement Learning: an Introduction*. 2nd ed. http://incompleteideas.net/book/the-book-2nd.html. MIT Press, 2018 (cit. on p. 4).

[46] C. Szepesvári and T. Lattimore. *Bandit Algorithms*. Cambridge University press, 2019 (cit. on p. 4).