RESEARCH CENTRE

Inria Center at the University of Lille

IN PARTNERSHIP WITH:

CNRS, Université de Lille

2022

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, Researcher]
- Debabrota Basu [INRIA, Researcher]
- Remy Degenne [INRIA, Researcher]
- Emilie Kaufmann [CNRS, Researcher, HDR]
- Odalric-Ambrym Maillard [INRIA, Researcher, HDR]

Faculty Member

- Philippe Preux [Team leader, UNIV LILLE, Professor, HDR]

Post-Doctoral Fellows

- Tuan Dam Quang Tuan [inria, from Oct 2022]
- Riccardo Della Vecchia [INRIA]
- Timothee Mathieu [INRIA]
- Mohit Mittal [INRIA, until Sep 2022]
- Alena Shilova [INRIA]
- Eduardo Vasconcellos [FFU - NITEROI, from Jul 2022]

PhD Students

- Achraf Azize [UNIV LILLE]
- Dorian Baudry [CNRS]
- Romain Gautron [cirad et CGIAR]
- Nathan Grinsztajn [LIX]
- Marc Jourdan [UNIV LILLE]
- Hector Kohler [INRIA, from Mar 2022 until Aug 2022]
- Penanklihi Cyrille Kone [INRIA, from Apr 2022 until Aug 2022]
- Matheus Medeiros Centa [UNIV LILLE]
- Reda Ouhamma [UNIV LILLE, until Aug 2022]
- Fabien Pesquerel [ENS PARIS]
- Patrick Saux [INRIA]
- Sumit Vashishtha [UNIV LILLE, from Oct 2022]
- omar darwiche [Inria]
- johan ferret [Google]
- romain gautron [Cirad/CGIAR]
- julien tarbouriech [facebook]

Technical Staff

- Hernan David Carvajal Bastidas [INRIA, Engineer]
- Tomy Soumphonphakdy [inria, Engineer, from May 2022]

Administrative Assistants

- Aurore Dalle [INRIA]
- Lucille Leclercq [inria, until Aug 2022]
- Anne Rejl [INRIA]
- Amélie Supervielle [inria]

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 [53] and the reinforcement learning (RL) problem [52]. In a simplified way, RL considers the problem of learning an optimal policy in a Markov Decision Problem (MDP) [50]; 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 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 represents 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 interesting by RL (see e.g. [51]).

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 investigate patient follow-up with Prof. F. Pattou's research group (CHU Lille, Inserm, Université de Lille) in project B4H. This effort comes along with investigating how we may use medical data available locally at CHU Lille, and also the national social security data. We also investigate drug repurposing with Prof. A. Delahaye-Duriez (Inserm, Université de Paris) in project Repos. We also study 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

We submitted two ANR JCJC and one ANR PRC projects and all 3 were accepted. They begin in 2023 and will last 4 years.

BIP-UP is an ANR PRC with Inserm U. 1190, headed by Ph. Preux.

FATE is an ANR JCJC headed by R. Degenne.

REPUBLIC is an ANR JCJC headed by D. Basu.

6.1 Awards

- D. Basu and collaborators received the Best Paper with Student Presenter Award in ACM EAAMO 2022 for their paper “On Meritocracy in Optimal Set Selection".
- T. Mathieu received the “prix solennel de thèse” of the Chancellerie des Universités de Paris in 2022.

7 New software and platforms

7.1 New software

7.1.1 rlberry

Keywords: Reinforcement learning, Simulation, Artificial intelligence

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 gym-DSSAT

Keywords: Reinforcement learning, Crop management, Sequential decision making under uncertainty, Mechanistic modeling

Functional Description: gym-DSSAT let you (learn to) manage a crop parcel, from seed selection, to daily activity in the field, to harvesting.

URL: https://gitlab.inria.fr/rgautron/gym_dssat_pdi

Contact: Romain Gautron

Partners: CIRAD, Cgiar

7.1.3 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 rench 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

8 New results

We organize our research results in a set of categories. The main categories are: bandit problems, reinforcement learning problems, and applications.

Participants: all Scool members.

8.1 Bandits and RL theory

Efficient Algorithms for Extreme Bandits, [17]

In this paper, we contribute to the Extreme Bandit problem, a variant of Multi-Armed Bandits in which the learner seeks to collect the largest possible reward. We first study the concentration of the maximum of i.i.d random variables under mild assumptions on the tail of the rewards distributions. This analysis motivates the introduction of Quantile of Maxima (QoMax). The properties of QoMax are sufficient to build an Explore-Then-Commit (ETC) strategy, QoMax-ETC, achieving strong asymptotic guarantees despite its simplicity. We then propose and analyze a more adaptive, anytime algorithm, QoMax-SDA, which combines QoMax with a subsampling method recently introduced by Baudry et al. (2021). Both algorithms are more efficient than existing approaches in two aspects (1) they lead to better empirical performance (2) they enjoy a significant reduction of the memory and time complexities.

Optimistic PAC Reinforcement Learning: the Instance-Dependent View, [37]

Optimistic algorithms have been extensively studied for regret minimization in episodic tabular MDPs, both from a minimax and an instance-dependent view. However, for the PAC RL problem, where the goal is to identify a near-optimal policy with high probability, little is known about their instance-dependent sample complexity. A negative result of Wagenmaker et al. (2022) suggests that optimistic sampling rules cannot be used to attain the (still elusive) optimal instance-dependent sample complexity. On the positive side, we provide the first instance-dependent bound for an optimistic algorithm for PAC RL, BPI-UCRL, for which only minimax guarantees were available (Kaufmann et al., 2021). While our bound features some minimal visitation probabilities, it also features a refined notion of sub-optimality gap compared to the value gaps that appear in prior work. Moreover, in MDPs with deterministic transitions, we show that BPI-UCRL is actually near-optimal. On the technical side, our analysis is very simple thanks to a new "target trick" of independent interest. We complement these findings with a novel hardness result explaining why the instance-dependent complexity of PAC RL cannot be easily related to that of regret minimization, unlike in the minimax regime.

Near Instance-Optimal PAC Reinforcement Learning for Deterministic MDPs, [31]

In probably approximately correct (PAC) reinforcement learning (RL), an agent is required to identify an ϵ-optimal policy with probability $1-\delta$. While minimax optimal algorithms exist for this problem, its instance-dependent complexity remains elusive in episodic Markov decision processes (MDPs). In this paper, we propose the first nearly matching (up to a horizon squared factor and logarithmic terms) upper and lower bounds on the sample complexity of PAC RL in deterministic episodic MDPs with finite state and action spaces. In particular, our bounds feature a new notion of sub-optimality gap for state-action pairs that we call the deterministic return gap. While our instance-dependent lower bound is written as a linear program, our algorithms are very simple and do not require solving such an optimization problem during learning. Their design and analyses employ novel ideas, including graph-theoretical concepts (minimum flows) and a new maximum-coverage exploration strategy.

Choosing Answers in epsilon-Best-Answer Identification for Linear Bandits, [23]

In pure-exploration problems, information is gathered sequentially to answer a question on the stochastic environment. While best-arm identification for linear bandits has been extensively studied in recent years, few works have been dedicated to identifying one arm that is ε-close to the best one (and not exactly the best one). In this problem with several correct answers, an identification algorithm should focus on one candidate among those answers and verify that it is correct. We demonstrate that picking the answer with highest mean does not allow an algorithm to reach asymptotic optimality in terms of expected sample complexity. Instead, a *furthest answer* should be identified. Using that insight to choose the candidate answer carefully, we develop a simple procedure to adapt best-arm identification algorithms to tackle ε-best-answer identification in transductive linear stochastic bandits. Finally, we propose an asymptotically optimal algorithm for this setting, which is shown to achieve competitive empirical performance against existing modified best-arm identification algorithms.

Top Two Algorithms Revisited, [24]

Top Two algorithms arose as an adaptation of Thompson sampling to best arm identification in multi-armed bandit models [38], for parametric families of arms. They select the next arm to sample from by randomizing among two candidate arms, a leader and a challenger. Despite their good empirical performance, theoretical guarantees for fixed-confidence best arm identification have only been obtained

when the arms are Gaussian with known variances. In this paper, we provide a general analysis of Top Two methods, which identifies desirable properties of the leader, the challenger, and the (possibly non-parametric) distributions of the arms. As a result, we obtain theoretically supported Top Two algorithms for best arm identification with bounded distributions. Our proof method demonstrates in particular that the sampling step used to select the leader inherited from Thompson sampling can be replaced by other choices, like selecting the empirical best arm.

IMED-RL: Regret optimal learning of ergodic Markov decision processes, [26]

We consider reinforcement learning in a discrete, undiscounted, infinite-horizon Markov Decision Problem (MDP) under the average reward criterion, and focus on the minimization of the regret with respect to an optimal policy, when the learner does not know the rewards nor the transitions of the MDP. In light of their success at regret minimization in multi-armed bandits, popular bandit strategies, such as the optimistic UCB, KL-UCB or the Bayesian Thompson sampling strategy, have been extended to the MDP setup. Despite some key successes, existing strategies for solving this problem either fail to be provably asymptotically optimal, or suffer from prohibitive burn-in phase and computational complexity when implemented in practice. In this work, we shed a novel light on regret minimization strategies, by extending to reinforcement learning the computationally appealing Indexed Minimum Empirical Divergence (IMED) bandit algorithm. Traditional asymptotic problem-dependent lower bounds on the regret are known under the assumption that the MDP is ergodic. Under this assumption, we introduce IMED-RL and prove that its regret upper bound asymptotically matches the regret lower bound. We discuss both the case when the supports of transitions are unknown, and the more informative but a priori harder-to-exploit-optimally case when they are known. Rewards are assumed light-tailed, semi-bounded from above. Last, we provide numerical illustrations on classical tabular MDPs, ergodic and communicating only, showing the competitiveness of IMED-RL in finite-time against state-of-the-art algorithms. IMED-RL also benefits from a light complexity.

Risk-aware linear bandits with convex loss, [36]

In decision-making problems such as the multi-armed bandit, an agent learns sequentially by optimizing a certain feedback. While the mean reward criterion has been extensively studied, other measures that reflect an aversion to adverse outcomes, such as mean-variance or conditional value-at-risk (CVaR), can be of interest for critical applications (healthcare, agriculture). Algorithms have been proposed for such risk-aware measures under bandit feedback without contextual information. In this work, we study contextual bandits where such risk measures can be elicited as linear functions of the contexts through the minimization of a convex loss. A typical example that fits within this framework is the expectile measure, which is obtained as the solution of an asymmetric least-square problem. Using the method of mixtures for supermartingales, we derive confidence sequences for the estimation of such risk measures. We then propose an optimistic UCB algorithm to learn optimal risk-aware actions, with regret guarantees similar to those of generalized linear bandits. This approach requires solving a convex problem at each round of the algorithm, which we can relax by allowing only approximated solution obtained by online gradient descent, at the cost of slightly higher regret. We conclude by evaluating the resulting algorithms on numerical experiments.

Bilinear Exponential Family of MDPs: Frequentist Regret Bound with Tractable Exploration & Planning, [35]

We study the problem of episodic reinforcement learning in continuous stateaction spaces with unknown rewards and transitions. Specifically, we consider the setting where the rewards and transitions are modeled using parametric bilinear exponential families. We propose an algorithm, BEF-RLSVI, that a) uses penalized maximum likelihood estimators to learn the unknown parameters, b) injects a calibrated Gaussian noise in the parameter of rewards to ensure exploration, and c) leverages linearity of the exponential family with respect to an underlying RKHS to perform tractable planning. We further provide a frequentist regret analysis of BEF-RLSVI that yields an upper bound of $\tilde{O}((d^3H^3K)^{1/2})$, where d is the dimension of the parameters, H is the episode length, and K is the number of episodes. Our analysis improves the existing bounds for the bilinear exponential family of MDPs by $\sqrt{H}$ and removes the handcrafted clipping deployed in existing RLSVI-type algorithms. Our regret bound is order-optimal with respect to H and K.

8.2 Bandits and RL face Real-life constraints

Efficient Change-Point Detection for Tackling Piecewise-Stationary Bandits, [11]

We introduce GLR-klUCB, a novel algorithm for the piecewise iid non-stationary bandit problem with bounded rewards. This algorithm combines an efficient bandit algorithm, kl-UCB, with an efficient, parameter-free, changepoint detector, the Bernoulli Generalized Likelihood Ratio Test, for which we provide new theoretical guarantees of independent interest. Unlike previous non-stationary bandit algorithms using a change-point detector, GLR-klUCB does not need to be calibrated based on prior knowledge on the arms' means. We prove that this algorithm can attain a $O(\sqrt{TA\Upsilon_T \log(T)})$ regret in T rounds on some "easy" instances, where A is the number of arms and Υ_T the number of change-points, without prior knowledge of Υ_T. In contrast with recently proposed algorithms that are agnostic to Υ_T, we perform a numerical study showing that GLR-klUCB is also very efficient in practice, beyond easy instances.

Near-Optimal Collaborative Learning in Bandits, [27]

This paper introduces a general multi-agent bandit model in which each agent is facing a finite set of arms and may communicate with other agents through a central controller in order to identify-in pure exploration-or play-in regret minimizationits optimal arm. The twist is that the optimal arm for each agent is the arm with largest expected mixed reward, where the mixed reward of an arm is a weighted sum of the rewards of this arm for all agents. This makes communication between agents often necessary. This general setting allows to recover and extend several recent models for collaborative bandit learning, including the recently proposed federated learning with personalization [30]. In this paper, we provide new lower bounds on the sample complexity of pure exploration and on the regret. We then propose a near-optimal algorithm for pure exploration. This algorithm is based on phased elimination with two novel ingredients: a data-dependent sampling scheme within each phase, aimed at matching a relaxation of the lower bound.

Exploration in Reinforcement Learning: Beyond Finite State-Spaces, [39]

Reinforcement learning (RL) is a powerful machine learning framework to design algorithms that learn to make decisions and to interact with the world. Algorithms for RL can be classified as offline or online. In the offline case, the algorithm is given a fixed dataset, based on which it needs to compute a good decision-making strategy. In the online case, an agent needs to efficiently collect data by itself, by interacting with the environment: that is the problem of exploration in reinforcement learning. This thesis presents theoretical and practical contributions to online RL. We investigate the worst-case performance of online RL algorithms in finite environments, that is, those that can be modeled with a finite amount of states, and where the set of actions that can be taken by an agent is also finite. Such performance degrades as the number of states increases, whereas in real-world applications the state set can be arbitrarily large or continuous. To tackle this issue, we propose kernel-based algorithms for exploration that can be implemented for general state spaces, and for which we provide theoretical results under weak assumptions on the environment. Those algorithms rely on a kernel function that measures the similarity between different states, which can be defined on arbitrary state-spaces, including discrete sets and Euclidean spaces, for instance. Additionally, we show that our kernel-based algorithms are able to handle non-stationary environments by using time-dependent kernel functions, and we propose and analyze approximate versions of our methods to reduce their computational complexity. Finally, we introduce a scalable approximation of our kernel-based methods, that can be implemented with deep reinforcement learning and integrate different representation learning methods to define a kernel function.

Bilinear Exponential Family of MDPs: Frequentist Regret Bound with Tractable Exploration & Planning, [35]

We study the problem of episodic reinforcement learning in continuous stateaction spaces with unknown rewards and transitions. Specifically, we consider the setting where the rewards and transitions are modeled using parametric bilinear exponential families. We propose an algorithm, BEF-RLSVI, that a) uses penalized maximum likelihood estimators to learn the unknown parameters, b) injects a calibrated Gaussian noise in the parameter of rewards to ensure exploration, and c) leverages linearity of the exponential family with respect to an underlying RKHS to perform tractable planning. We further provide a frequentist regret analysis of BEF-RLSVI that yields an upper bound of $\tilde{O}((d^3H^3K)^{1/2})$, where

d is the dimension of the parameters, H is the episode length, and K is the number of episodes. Our analysis improves the existing bounds for the bilinear exponential family of MDPs by $\sqrt{H}$ and removes the handcrafted clipping deployed in existing RLSVI-type algorithms. Our regret bound is order-optimal with respect to H and K.

Risk-Sensitive Bayesian Games for Multi-Agent Reinforcement Learning under Policy Uncertainty, [34]

In stochastic games with incomplete information, the uncertainty is evoked by the lack of knowledge about a player's own and the other players' types, i.e. the utility function and the policy space, and also the inherent stochasticity of different players' interactions. In existing literature, the risk in stochastic games has been studied in terms of the inherent uncertainty evoked by the variability of transitions and actions. In this work, we instead focus on the risk associated with the *uncertainty over types*. We contrast this with the multi-agent reinforcement learning framework where the other agents have fixed stationary policies and investigate risk-sensitiveness due to the uncertainty about the other agents' adaptive policies. We propose risk-sensitive versions of existing algorithms proposed for risk-neutral stochastic games, such as Iterated Best Response (IBR), Fictitious Play (FP) and a general multi-objective gradient approach using dual ascent (DAPG). Our experimental analysis shows that risk-sensitive DAPG performs better than competing algorithms for both social welfare and general-sum stochastic games.

SENTINEL: Taming Uncertainty with Ensemble-based Distributional Reinforcement Learning, [20]

In this paper, we consider risk-sensitive sequential decision-making in Reinforcement Learning (RL). Our contributions are two-fold. First, we introduce a novel and coherent quantification of risk, namely composite risk, which quantifies the joint effect of aleatory and epistemic risk during the learning process. Existing works considered either aleatory or epistemic risk individually, or as an additive combination. We prove that the additive formulation is a particular case of the composite risk when the epistemic risk measure is replaced with expectation. Thus, the composite risk is more sensitive to both aleatory and epistemic uncertainty than the individual and additive formulations. We also propose an algorithm, SENTINEL-K, based on ensemble bootstrapping and distributional RL for representing epistemic and aleatory uncertainty respectively. The ensemble of K learners uses Follow The Regularised Leader (FTRL) to aggregate the return distributions and obtain the composite risk. We experimentally verify that SENTINEL-K estimates the return distribution better, and while used with composite risk estimates, demonstrates higher risk-sensitive performance than state-of-the-art risk-sensitive and distributional RL algorithms.

On Meritocracy in Optimal Set Selection, [19]

Typically, merit is defined with respect to some intrinsic measure of worth. We instead consider a setting where an individual's worth is relative: when a decision maker (DM) selects a set of individuals from a population to maximise expected utility, it is natural to consider the expected marginal contribution (EMC) of each person to the utility. We show that this notion satisfies an axiomatic definition of fairness for this setting. We also show that for certain policy structures, this notion of fairness is aligned with maximising expected utility, while for linear utility functions it is identical to the Shapley value. However, for certain natural policies, such as those that select individuals with a specific set of attributes (e.g. high enough test scores for college admissions), there is a trade-off between meritocracy and utility maximisation. We analyse the effect of constraints on the policy on both utility and fairness in an extensive experiments based on college admissions and outcomes in Norwegian universities.

How Biased is Your Feature?: Computing Fairness Influence Functions with Global Sensitivity Analysis, [47]

Fairness in machine learning has attained significant focus due to the widespread application of machine learning in high-stake decision-making tasks. Unless regulated with a fairness objective, machine learning classifiers might demonstrate unfairness/bias towards certain demographic populations in the data. Thus, the quantification and mitigation of the bias induced by classifiers have become a central concern. In this paper, we aim to quantify the influence of different features on the bias of a classifier. To this end, we propose a framework of Fairness Influence Function (FIF), and compute it as a scaled difference of conditional variances in the prediction of the classifier. We also instantiate an algorithm, FairXplainer, that uses variance decomposition among the subset of features and a local regressor to compute FIFs accurately, while also capturing the intersectional effects of the features. Our experimental analysis validates that FairXplainer captures the influences of both individual features and higher-order

feature interactions, estimates the bias more accurately than existing local explanation methods, and detects the increase/decrease in bias due to affirmative/punitive actions in the classifier.

Algorithmic fairness verification with graphical models, [22]

In recent years, machine learning (ML) algorithms have been deployed in safety-critical and high-stake decision-making, where the fairness of algorithms is of paramount importance. Fairness in ML centers on detecting bias towards certain demographic populations induced by an ML classifier and proposes algorithmic solutions to mitigate the bias with respect to different fairness definitions. To this end, several fairness verifiers have been proposed that compute the bias in the prediction of an ML classifier-essentially beyond a finite dataset-given the probability distribution of input features. In the context of verifying linear classifiers, existing fairness verifiers are limited by accuracy due to imprecise modeling of correlations among features and scalability due to restrictive formulations of the classifiers as SSAT/SMT formulas or by sampling. In this paper, we propose an efficient fairness verifier, called FVGM, that encodes the correlations among features as a Bayesian network. In contrast to existing verifiers, FVGM proposes a stochastic subset-sum based approach for verifying linear classifiers. Experimentally, we show that FVGM leads to an accurate and scalable assessment for more diverse families of fairness-enhancing algorithms, fairness attacks, and group/causal fairness metrics than the state-of-the-art. We also demonstrate that FVGM facilitates the computation of fairness influence functions as a stepping stone to detect the source of bias induced by subsets of features.

When Privacy Meets Partial Information: A Refined Analysis of Differentially Private Bandits, [16]

We study the problem of multi-armed bandits with ϵ-global Differential Privacy (DP). First, we prove the minimax and problem-dependent regret lower bounds for stochastic and linear bandits that quantify the hardness of bandits with ϵ-global DP. These bounds suggest the existence of two hardness regimes depending on the privacy budget ϵ. In the high-privacy regime (small ϵ), the hardness depends on a coupled effect of privacy and partial information about the reward distributions. In the low-privacy regime (large ϵ), bandits with ϵ-global DP are not harder than the bandits without privacy. For stochastic bandits, we further propose a generic framework to design a near-optimal ϵ global DP extension of an index-based optimistic bandit algorithm. The framework consists of three ingredients: the Laplace mechanism, arm-dependent adaptive episodes, and usage of only the rewards collected in the last episode for computing private statistics. Specifically, we instantiate ϵ-global DP extensions of UCB and KL-UCB algorithms, namely AdaP-UCB and AdaP-KLUCB. AdaP-KLUCB is the first algorithm that both satisfies ϵ-global DP and yields a regret upper bound that matches the problem-dependent lower bound up to multiplicative constants.

Procrastinated Tree Search: Black-box Optimization with Delayed, Noisy, and Multi-fidelity Feedback, [32]

In black-box optimization problems, we aim to maximize an unknown objective function, where the function is only accessible through feedbacks of an evaluation or simulation oracle. In real-life, the feedbacks of such oracles are often noisy and available after some unknown delay that may depend on the computation time of the oracle. Additionally, if the exact evaluations are expensive but coarse approximations are available at a lower cost, the feedbacks can have multi-fidelity. In order to address this problem, we propose a generic extension of hierarchical optimistic tree search (HOO), called ProCrastinated Tree Search (PCTS), that flexibly accommodates a delay and noise-tolerant bandit algorithm. We provide a generic proof technique to quantify regret of PCTS under delayed, noisy, and multi-fidelity feedbacks. Specifically, we derive regret bounds of PCTS enabled with delayed-UCB1 (DUCB1) and delayed-UCB-V (DUCBV) algorithms. Given a horizon T, PCTS retains the regret bound of non-delayed HOO for expected delay of $O(logT)$, and worsens by $T^{(1-\alpha)/(d+2)}$ for expected delays of $O(T^{1-\alpha})$ for $\alpha \in (0,1]$. We experimentally validate on multiple synthetic functions and hyperparameter tuning problems that PCTS outperforms the state-of-the-art black-box optimization methods for feedbacks with different noise levels, delays, and fidelity.

UDO: Universal Database Optimization using Reinforcement Learning, [33]

UDO is a versatile tool for offline tuning of database systems for specific workloads. UDO can consider a variety of tuning choices, reaching from picking transaction code variants over index selections up to database system parameter tuning. UDO uses reinforcement learning to converge to near-optimal configurations, creating and evaluating different configurations via actual query executions (instead of relying on simplifying cost models). To cater to different parameter types, UDO distinguishes heavy

parameters (which are expensive to change, e.g. physical design parameters) from light parameters. Specifically for optimizing heavy parameters, UDO uses reinforcement learning algorithms that allow delaying the point at which the reward feedback becomes available. This gives us the freedom to optimize the point in time and the order in which different configurations are created and evaluated (by benchmarking a workload sample). UDO uses a cost-based planner to minimize reconfiguration overheads. For instance, it aims to amortize the creation of expensive data structures by consecutively evaluating configurations using them. We evaluate UDO on Postgres as well as MySQL and on TPC-H as well as TPC-C, optimizing a variety of light and heavy parameters concurrently.

Bandits Corrupted by Nature: Lower Bounds on Regret and Robust Optimistic Algorithm, [41]

In this paper, we study the stochastic bandits problem with k unknown heavy-tailed and corrupted reward distributions or arms with time-invariant corruption distributions. At each iteration, the player chooses an arm. Given the arm, the environment returns an uncorrupted reward with probability $1-\epsilon$ and an arbitrarily corrupted reward with probability ϵ. In our setting, the uncorrupted reward might be heavy-tailed and the corrupted reward might be unbounded. We prove a lower bound on the regret indicating that the corrupted and heavy-tailed bandits are strictly harder than uncorrupted or light-tailed bandits. We observe that the environments can be categorised into hardness regimes depending on the suboptimality gap Δ, variance σ, and corruption proportion ϵ. Following this, we design a UCB-type algorithm, namely HuberUCB, that leverages Huber's estimator for robust mean estimation. HuberUCB leads to tight upper bounds on regret in the proposed corrupted and heavy-tailed setting. To derive the upper bound, we prove a novel concentration inequality for Huber's estimator, which might be of independent interest.

SAAC: Safe Reinforcement Learning as an Adversarial Game of Actor-Critics, [21]

Although Reinforcement Learning (RL) is effective for sequential decision-making problems under uncertainty, it still fails to thrive in real-world systems where risk or safety is a binding constraint. In this paper, we formulate the RL problem with safety constraints as a non-zero-sum game. While deployed with maximum entropy RL, this formulation leads to a safe adversarially guided soft actor-critic framework, called SAAC. In SAAC, the adversary aims to break the safety constraint while the RL agent aims to maximize the constrained value function given the adversary's policy. The safety constraint on the agent's value function manifests only as a repulsion term between the agent's and the adversary's policies. Unlike previous approaches, SAAC can address different safety criteria such as safe exploration, mean-variance risk sensitivity, and CVaR-like coherent risk sensitivity. We illustrate the design of the adversary for these constraints. Then, in each of these variations, we show the agent differentiates itself from the adversary's unsafe actions in addition to learning to solve the task. Finally, for challenging continuous control tasks, we demonstrate that SAAC achieves faster convergence, better efficiency, and fewer failures to satisfy the safety constraints than risk-averse distributional RL and risk-neutral soft actor-critic algorithms.

Online Instrumental Variable Regression: Regret Analysis and Bandit Feedback, [44]

The independence of noise and covariates is a standard assumption in online linear regression and linear bandit literature. This assumption and the following analysis are invalid in the case of endogeneity, i.e., when the noise and covariates are correlated. In this paper, we study the online setting of instrumental variable (IV) regression, which is widely used in economics to tackle endogeneity. Specifically, we analyse and upper bound regret of Two-Stage Least Squares (2SLS) approach to IV regression in the online setting. Our analysis shows that Online 2SLS (O2SLS) achieves $O(d^2 \log^2 T)$ regret after T interactions, where d is the dimension of covariates. Following that, we leverage the O2SLS as an oracle to design OFUL-IV, a linear bandit algorithm. OFUL-IV can tackle endogeneity and achieves $O(d\sqrt{T}\log T)$ regret. For datasets with endogeneity, we experimentally demonstrate that O2SLS and OFUL-IV incur lower regrets than the state-of-the-art algorithms for both the online linear regression and linear bandit settings.

IMED-RL: Regret optimal learning of ergodic Markov decision processes, [26]

We consider reinforcement learning in a discrete, undiscounted, infinite-horizon Markov Decision Problem (MDP) under the average reward criterion, and focus on the minimization of the regret with respect to an optimal policy, when the learner does not know the rewards nor the transitions of the MDP. In light of their success at regret minimization in multi-armed bandits, popular bandit strategies, such as the optimistic UCB, KL-UCB or the Bayesian Thompson sampling strategy, have been extended to the MDP setup. Despite some key successes, existing strategies for solving this problem either fail to be provably asymptotically optimal, or suffer from prohibitive burn-in phase and computational complexity

when implemented in practice. In this work, we shed a novel light on regret minimization strategies, by extending to reinforcement learning the computationally appealing Indexed Minimum Empirical Divergence (IMED) bandit algorithm. Traditional asymptotic problem-dependent lower bounds on the regret are known under the assumption that the MDP is ergodic. Under this assumption, we introduce IMED-RL and prove that its regret upper bound asymptotically matches the regret lower bound. We discuss both the case when the supports of transitions are unknown, and the more informative but a priori harder-to-exploit-optimally case when they are known. Rewards are assumed light-tailed, semi-bounded from above. Last, we provide numerical illustrations on classical tabular MDPs, ergodic and communicating only, showing the competitiveness of IMED-RL in finite-time against state-of-the-art algorithms. IMED-RL also benefits from a light complexity.

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

Entropy Regularized Reinforcement Learning with Cascading Networks, [45]

Deep Reinforcement Learning (Deep RL) has had incredible achievements on high dimensional problems, yet its learning process remains unstable even on the simplest tasks. Deep RL uses neural networks as function approximators. These neural models are largely inspired by developments in the (un)supervised machine learning community. Compared to these learning frameworks, one of the major difficulties of RL is the absence of i.i.d. data. One way to cope with this difficulty is to control the rate of change of the policy at every iteration. In this work, wechallenge the common practices of the (un)supervised learning community of using a fixed neural architecture, by having a neural model that grows in size at each policy update. This allows a closed form entropy regularized policy update, which leads to a better control of the rate of change of the policy at each iteration and help cope with the non i.i.d. nature of RL. Initial experiments on classical RL benchmarks show promising results with remarkable convergence on some RL tasks when compared to other deep RL baselines, while exhibiting limitations on others.

Automated planning for robotic guidewire navigation in the coronary arteries, [29]

Soft continuum robots, and comparable instruments allow to perform some surgical procedures noninvasively. While safer, less morbid and more cost-effective, these medical interventions increase the complexity for the practitioners: the manipulation of anatomical structures is indirect through telescopic and flexible devices and the visual feedback is indirect through monitors. Interventional cardiology is an example of complex procedures where catheters and guidewires are manipulated to reach and treat remote areas of the vascular network. Such interventions may be assisted with a robot that will operate the tools but the planning (choice of tools and trajectories) remains a complex task. In this paper we use a simulation framework for flexible devices inside the vasculature and we propose a method to automatically control these devices to reach specific locations. Experiments performed on 15 patient geometries exhibit good performance. Automatic manipulation reaches the goal in more than 90% of the cases.

SofaGym: An open platform for Reinforcement Learning based on Soft Robot simulations, [15]

OpenAI Gym is one of the standard interfaces used to train Reinforcement Learning (RL) Algorithms. The Simulation Open Framework Architecture (SOFA) is a physics based engine that is used for soft robotics simulation and control based on real-time models of deformation. The aim of this paper is to present SofaGym, an open source software to create OpenAI Gym interfaces, called environments, out of soft robot digital twins.The link between soft robotics and RL offers new challenges for both fields: representation of the soft robot in a RL context, complex interactions with the environment, use of specific mechanical tools to control soft robots, transfer of policies learned in simulation to the real world, etc. The article presents the large possible uses of SofaGym to tackle these challenges by using RL and planning algorithms. This publication contains neither new algorithms nor new models but proposes a new platform, open to the community, that offers non existing possibilities of coupling RL to physics based simulation of soft robots. We present 11 environments, representing a wide variety of soft robots and applications, we highlight the challenges showcased by each environment. We propose methods of solving the task using traditional control, RL and planning and point out research perspectives using the platform.

Reinforcement Learning for crop management, [13]

Reinforcement Learning (RL), including Multi-Armed Bandits, is a branch of Machine Learning that deals with the problem of sequential decision-making in uncertain and unknown environments through learning by practice. While best known for being the core of the Artificial Intelligence (AI) world's best Go game player, RL has a vast potential range of applications. RL may help to address some of the criticisms leveled against crop management Decision Support Systems (DSS): it is an interactive, geared toward action, contextual tool to evaluate series of crop operations faced with uncertainties. A review of RL use for crop management DSS reveals a limited number of contributions. We profile key prospects for a human-centered, real world, interactive RL-based system to face tomorrow's agricultural decisions and theoretical and ongoing practical challenges that may explain its current low take-up. We argue that a joint research effort from the RL and agronomy communities is necessary to explore RL's full potential.

gym-DSSAT: a crop model turned into a Reinforcement Learning environment, [46]

Addressing a real world sequential decision problem with Reinforcement Learning (RL) usually starts with the use of a simulated environment that mimics real conditions. We present a novel open source RL environment for realistic crop management tasks. gym-DSSAT is agym interface to the Decision Support System for Agrotechnology Transfer (DSSAT), a high fidelity crop simulator. DSSAT has been developped over the last 30 years and is widely recognized by agronomists. gym-DSSAT comes with predefined simulations based on real world maize experiments.The environment is as easy to use as any gym environment. We provide performance baselines using basic RL algorithms. We also briefly outline how the monolithic DSSAT simulator written in Fortran has been turned into a Python RL environment. Our methodology is generic and may be applied to similar simulators. We report on very preliminary experimental results which suggest that RL can help researchers to improve sustainability of fertilization and irrigation practices.

Foundations and state of the art, [38]

In this chapter, we address the foundational aspects of digital technology, their use in agriculture and current research

Combination of gene regulatory networks and sequential machine learning for drug repurposing, [40]

Given the ever increasing cost of designing de novo molecules to target causes of diseases, and the huge amount of currently available biological data, the development of systematic explorative pipelines for drug development has become of paramount importance. In my thesis, I focused on drug repurposing, which is a paradigm that aims at identifying new therapeutic indications for known chemical compounds. Due to the already large collection of transcriptomic data -that is, related to protein production through the transcription of gene DNA sequences- which is publicly available, I investigated how to process in a transparent and controllable way this information about gene activity to screen molecules. The current state of research in drug development indicates that such generic approaches might considerably fasten the discovery of promising therapies, especially for neglected or rare diseases research. First, noting that transcriptomic measurements are the product of a complex dynamical system of co- and inter-gene activity regulations, I worked on integrating in an automated fashion diverse types of biological information in order to build a model of these regulations. That is where gene regulatory networks, and more specifically, Boolean networks, intervene. Such models are useful for both explaining observed transcription levels, and for predicting the result of gene activity perturbations through molecules. Second, these models allow online in silico drug testing. While using the predictive features of Boolean networks can be costly, the core assumption of this thesis is that, combining them with sequential learning algorithms, such as multi-armed bandits, might mitigate that effect, and help control the error rate in recommended therapeutic candidates. This is the drug testing procedure suggested throughout my PhD. The question of the proper integration of known side information about the chemical compounds into multi-armed bandits is crucial, and has also been investigated further. Finally, I applied part of my work to ranking different treatment protocols for neurorepair in the case of encephalopathy in premature infants. On the theoretical side, I also contributed to the design of an algorithm which is able to extend the drug testing procedure in a distributed way, for instance across several tested populations, disease models, or research teams.

An Integer Linear Programming Approach for Pipelined Model Parallelism, [42]

The training phase in Deep Neural Networks has become an important source of computing resource usage and because of the resulting volume of computation, it is crucial to perform it efficiently on

parallel architectures. Even today, data parallelism is the most widely used method, but the associated requirement to replicate all the weights on the totality of computation resources poses problems of memory at the level of each node and of collective communications at the level of the platform. In this context, the model parallelism, which consists in distributing the different layers of the network over the computing nodes, is an attractive alternative. Indeed, it is expected to better distribute weights (to cope with memory problems) and it does not imply large collective communications since only forward activations are communicated. However, to be efficient, it must be combined with a pipelined/streaming approach, which leads in turn to new memory costs. The goal of this paper is to model these memory costs in detail and to show that it is possible to formalize this optimization problem as an Integer Linear Program (ILP).

MadPipe: Memory Aware Dynamic Programming Algorithm for Pipelined Model Parallelism, [18]

The training phase in Deep Neural Networks (DNNs) is very computationally intensive and is nowadays often performed on parallel computing platforms, ranging from a few GPUs to several thousand GPUs. The strategy of choice for the parallelization of training is the so-called data parallel approach, based of the parallel training of the different inputs (typically images) and a the aggregation of network weights with collective communications (AllReduce). The scalability of this approach is limited both by the memory available on each node and the networking capacities for collective operations. Recently, a parallel model approach, in which the network weights are distributed and images are trained in a pipeline/stream manner over the computational nodes has been proposed (Pipedream, Gpipe). In this paper, we formalize in detail the optimization problem associated with the placement of DNN layers onto computation resources when using pipelined model parallelism, and we derive a dynamic programming based heuristic, MadPipe, that allows to significantly improve the performance of the parallel model approach compared to the literature.

Weight Offloading Strategies for Training Large DNN Models, [43]

The limited memory of GPUs induces serious problems in the training phase of deep neural networks (DNNs). Indeed, with the recent tremendous increase in the size of DNN models, which can now routinely include hundreds of billions or even trillions of parameters, it is impossible to store these models in the memory of a GPU and several strategies have been devised to solve this problem. In this paper, we analyze in detail the strategy that consists in offloading the weights of some model layers from the GPU to the CPU when they are not used. Since the PCI bus bandwidth between the GPU and the CPU is limited, it is crucial to know which layers should be transferred (offloaded and prefetched) and when. We prove that this problem is in general NP-Complete in the strong sense and we propose a lower bound formulation in the form of an Integer Linear Program (ILP). We propose heuristics to select the layers to offload and to build the schedule of data transfers. We show that this approach allows to build near-optimal weight offloading strategies on realistic size DNNs and architectures.

8.4 Other

Topics in robust statistical learning, [12]

Some recent contributions to robust inference are presented. Firstly, the classical problem of robust M-estimation of a location parameter is revisited using an optimal transport approach-with specifically designed Wasserstein-type distances-that reduces robustness to a continuity property. Secondly, a procedure of estimation of the distance function to a compact set is described, using union of balls. This methodology originates in the field of topological inference and offers as a byproduct a robust clustering method. Thirdly, a robust Lloyd-type algorithm for clustering is constructed, using a bootstrap variant of the median-of-means strategy. This algorithm comes with a robust initialization.

Concentration study of M-estimators using the influence function, [14]

We present a new finite-sample analysis of M-estimators of locations in a Hilbert space using the tool of the influence function. In particular, we show that the deviations of an M-estimator can be controlled thanks to its influence function (or its score function) and then, we use concentration inequality on M-estimators to investigate the robust estimation of the mean in high dimension in a corrupted setting (adversarial corruption setting) for bounded and unbounded score functions. For a sample of size n and covariance matrix Σ, we attain the minimax speed $Tr(\Sigma)/n + \Sigma op \log(1/\delta)/n$ with probability larger than $1 - \delta$ in a heavy-tailed setting. One of the major advantages of our approach compared to others

recently proposed is that our estimator is tractable and fast to compute even in very high dimension with a complexity of $O(ndlog(Tr(\Sigma)))$ where n is the sample size and Σ is the covariance matrix of the inliers and in the code that we make available for this article is tested to be very fast.

9 Bilateral contracts and grants with industry

9.1 Bilateral contracts with industry

Participants: Philippe Preux, Léonard Hussenot, Johan Ferret, Jean Tarbouriech.

- 2 contracts with Google regarding PhDs of J. Ferret and L. Hussenot (2020–2022), managed by Ph. Preux.
- 1 contract with Facebook AI Research regarding PhD of J. Tarbouriech (2020–2022), managed by Ph. Preux.

10 Partnerships and cooperations

Participants: all Scool permanent members.

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: Philippe Preux

Partners:

- Bihar Agriculture University, India,
- Inria FUN, Lille.

Inria contact: Philippe Preux

Summary: as part of our research activities related to the application of reinforcement learning and bandits to agriculture, this associate teams aim at providing us with in-field data, and also the ability to perform in-field experiments. This sort of experiments is extremely useful to train our algorithms which have to explore, that is test new actions in the field and observe their outcome. This approach is complementary to the one we investigate with the use of the DSSAT simulator.

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.

10.1.2 STIC AmSud projects

Title: emistral

Duration: 2021 → 2022

Coordinator: Luis Marti (Inria CHile)

Partners:

- Inria Chile
- UFF, Niteroi, Brazil
- Universidad de la República, Montevideo, Uruguay
- Inria Scool, Inria AIO

Inria contact: Philippe Preux

Summary: The current climate crisis calls for the use of all available technology to try to understand, model, predict and hopefully work towards its mitigation. Oceans play a key role in grasping the complex and intertwined processes that govern these phenomena. Oceans -and rivers- play a key role in regulating the planet's climate, weather and ecology. Recent advances in computer sciences and applied mathematics, such as machine learning, artificial intelligence, scientific computation, among others, have produced a revolution in our capacity for understanding the emergence of patterns and dynamics in complex systems while at the same time the complexity of these problems pose significant challenges to computer science itself. The key factor deciding about the success of failure of the application of these methods is having sufficient and adequate data. Oceanographic vessels have been extensively used to gather this data. However, they have been shown to be insufficient because their high operation cost, the risks involved and their limited availability. Autonomous sailboats present themselves as a viable alternative. In principle, by relying on wind energy they could operate for indefinite periods being only limited by the effects of fouling and the wear and tear of materials. Recent results in the area of machine learning are especially suited to fill this gap. In particular, reinforcement learning (RL), transfer learning (TL) and autonomous learning (AL). The combination of those methods could overcome the need of programming particular controller for every boat as it would be capable of replicating at some degree, the learning process of human skippers and sailors.

10.2 International research visitors

10.2.1 Visits of international scientists

Other international visits to the team

- Agrawal Shubhada, post-doc at GeorgiaTech, Apr-June 2022.

10.3 European initiatives

10.3.1 Other european programs/initiatives

Title: CausalXRL

Duration: 2021 → 2024

Coordinator: Aditya Gilra, U. Amsterdam

Partners:

- U. Amsterdam
- U. Sheffield
- U. Vienna
- Inria Scool

Inria contact: Philippe Preux

Summary: Deep reinforcement learning systems are approaching or surpassing human-level performance in specific domains, from games to decision support to continuous control, albeit in non-critical environments. Most of these systems require random exploration and state-action-value-based exploitation of the environment. However, in important real-life domains, like medical decision support or patient rehabilitation, every decision or action must be fully justified and certainly not random. We propose to develop neural networks that learn causal models of the environment relating action to effect, initially using offline data. The models will then be interfaced with reinforcement learning and decision support networks, so that every action taken online can be explained or justified based on its expected effect. The causal model can then be refined iteratively, enabling to better predict future cascading effects of any action chain. The system, subsequently termed CausalXRL, will only propose actions that can be justified on the basis of beneficial effects. When the immediate benefit is uncertain, the system will propose explorative actions that generate most-probable future benefit. CausalXRL thus supports the user in choosing actions based on specific expected outcomes, rather than as prescribed by a black box.

10.4 National initiatives

Scool is involved in 1 ANR project:

- ANR Bold, headed by V. Perchet (ENS Paris-Saclay, ENSAE), local head: É. Kaufmann, 2019–2023.

Scool is involved in some Inria projects:

- Challenge HPC – Big Data, headed by B. Raffin, Datamove, Grenoble.

  In this challenge, we collaborate with:

  - B. Raffin, on what HPC can bring and can be used at its best for reinforcement learning.
  - O. Beaumont, E. Jeannot, on what RL can bring to HPC, in particular the use of RL for task scheduling.

- Challenge HY_AIAI.

  In this challenge, we collaborate with L. Gallaraga, CR Inria Rennes, about the combination of statistical and symbolic approaches in machine learning.

- Exploratory action "Sequential Recommendation for Sustainable Gardening (SR4SG)", headed by O-A. Maillard.

Other collaborations in France:

- R. Gautron, PhD student, Cirad, agricultural practices recommendation.
- L. Soulier, Associate Professor, Sorbonne Université, reinforcement learning for information retrieval.
- M. Valko, researcher DeepMind.
- A. Delahaye-Duriez, INSERM, Université de Paris.
- B. De-Saporta, Université de Montpellier, piecewise-deterministic Markov processes.
- A. Garivier, Professor, ENS Lyon
- V. Perchet, Professor, ENSAE & Criteo AI Lab
- P. Gaillard, CR, Inria Grenoble - Rhône-Alpes
- A. Bellet, CR, Inria Lille-Nord Europe (Équipe Magnet)

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. 2020–2023.

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

11 Dissemination

Participants: many Scool members.

11.1 Promoting scientific activities

11.1.1 Scientific events: organisation

General chair, scientific chair

- R. Degenne co-organized the Complex Feedback in Online Learning workshop at ICML 2022.

11.1.2 Scientific events: selection

Member of the conference program committees

- E. Kaufmann is a PC member for ALT and EWRL.
- O-A. Maillard is a PC member for ICML, AISTATS.
- Ph. Preux is an SPC for AAAI, and PC for IJCAI and ECML.

Reviewer

- R. Degenne: ICML and AISTATS

11.1.3 Journal

Member of the editorial boards

- O-A. Maillard is in the editorial board of JMLR.

Reviewer - reviewing activities

- R. Akrour: JMLR
- R. Degenne: JMLR
- E. Kaufmann: IEEE IT, IEEE Transactions on Games, Statistica Sinica
- O-A. Maillard: Journal of Machine Learning Research (JMLR), Autonomous Agents and Multi-Agent Systems (AGNT), Machine Learning (MACH), The Annals of Statistics (AoS)

11.1.4 Invited talks

- R. Degenne: invited talk at the Probability & Statistics Group Heidelberg-Mannheim
- E. Kaufmann: plenary talk at the Journées de Statistiques, invited speaker at SNB (Statistics and Biopharmacy) 2022, invited talks at the Harvard Statistics COlloquium (virtual)

11.1.5 Scientific expertise

- Ph. Preux is:
  - a member of the IRD CSS 5 (data science and models),
  - a member of the Commission d'Évaluation (CE) of Inria,
  - a member of the scientific committee on ethics of the Health Data warehouse of the CHU de Lille.
- O-A. Maillard is:
  - a member of the Commission Emploi Recherche (CER) of Inria Lille.

11.1.6 Research administration

- Ph. Preux was deputy scientific delegate at Inria Lille until June 2022.

11.2 Teaching - Supervision - Juries

11.2.1 Teaching

- R. Akrour: Apprentissage à partir de données humaines, M1 in Cognitive Science, Université de Lille
- R. Akrour: Perception et motricité 2, L2 MIASHS, Université de Lille
- R. Akrour: Perception et motricité 1, L1 MIASHS, Université de Lille
- R. Degenne: Sequential learning, M2 MVA, ENS Paris-Saclay
- R. Degenne: Sequential learning, Centrale Lille
- R. Degenne: Sciences des données 3, L3 MIASHS, Université de Lille

- E. Kaufmann: Sequential Decision Making (24h), M2 Data Science, Ecole Centrale Lille.
- O-A. Maillard: Statistical Reinforcement Learning (48h), MAP/INF641, Master Artificial Intelligence and advanced Visual Computing, École Polytechnique.
- O-A. Maillard: Reinforcement Learning (24h), Master 2 Artificial Intelligence, École CentraleSupélec.
- Ph. Preux: « IA et apprentissage automatique », DU IA & Santé, Université de 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.

11.2.2 Supervision

- R. Akrour and Ph. Preux supervised the internship of:
  - Hector Kolher, M2 computer science, Sorbonne Université, Paris,
- R. Akrour and D. Basu supervised the internship of:
  - Mahdi Kallel, M2 Optimization, Institut Polytechnique de Paris, Paris,
- É. Kaufmann supervised the internship of:
  - Cyrille Kone, MVA.

11.2.3 Juries

- E. Kaufmann was a member of the juries of:
  - Ph.D. in CS of Léonard Blier, Université Paris-Saclay
  - Ph.D. in maths of Solenne Gaucher, Université Paris-Saclay
  - Ph.D. in CS of Geovani Rizk, Université Paris Dauphine
  - Ph.D. in CS of Arnaud Delaruyelle, Université de Lille
  - Ph.D. in maths of El Mehdi Saad, Université Paris-Saclay
  - Ph.D. in CS of Sarah Perrin, Université de Lille
- O-A. Maillard was a member of the juries of:
  - Ph.D. in Agronomy of Romain Gautron, Université de Montpellier
- Ph. Preux was a member of the juries of:
  - Ph.D. in CS of Camille-Sovanneary Gauthier, Université de Rennes
  - Ph.D. in CS of Pierre Schegg, Université de Lille
  - Ph.D. in CS of Johan Ferret, Université de Lille
  - Ph.D. in CS of Jean Tarbouriech, Université de Lille
  - Ph.D. in CS of Léonard Hussenot, Université de Lille
  - Ph.D. in Agronomy of Romain Gautron, Université de Montpellier
  - Ph.D. in CS of David Saltiel, Université du Littoral Côte d'Opale
  - Ph.D. in CS of Sigfried Delannoy, Université du Littoral Côte d'Opale

11.3 Popularization

11.3.1 Articles and contents

- an article on the Inria website regarding the Bandits for Health project.

11.3.2 Interventions

- T. Mathieu gave a talk on « Les statistiques ne servent pas qu'à nous espionner » (*statistics are not just spying on us*), Université D'Anchin, Douai, Oct. 2022.

11.3.3 Other mediation actions

- Ph. Preux is involved in the Merlin project at Université de Lille on "The big investigation on AI". The outcome of this project is a TV program produced by « L'esprit Sorcier TV », broadcasted in February 2023, and then available on replay.
- Ph. Preux is part of the scientific committee of the « Forum des Sciences » in Villeneuve d'Ascq regarding the season on AI.

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] L. Besson and E. Kaufmann. 'Multi-Player Bandits Revisited'. In: *Algorithmic Learning Theory*. Mehryar Mohri and Karthik Sridharan. Lanzarote, Spain, Apr. 2018. URL: https://hal.inria.fr/hal-01629733.

[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] A. Garivier and E. Kaufmann. 'Optimal Best Arm Identification with Fixed Confidence'. In: *29th Annual Conference on Learning Theory (COLT)*. Vol. 49. JMLR Workshop and Conference Proceedings. New York, United States, June 2016. URL: https://hal.archives-ouvertes.fr/hal-01273838.

[6] 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.

[7] E. Kaufmann and W. M. Koolen. 'Monte-Carlo Tree Search by Best Arm Identification'. In: *NIPS 2017 - 31st Annual Conference on Neural Information Processing Systems*. Advances in Neural Information Processing Systems. Long Beach, United States, Dec. 2017, pp. 1–23. URL: https://hal.archives-ouvertes.fr/hal-01535907.

[8] 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.

[9] 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.

[10] 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.

12.2 Publications of the year

International journals

[11] L. Besson, E. Kaufmann, O.-A. Maillard and J. Seznec. 'Efficient Change-Point Detection for Tackling Piecewise-Stationary Bandits'. In: *Journal of Machine Learning Research* (Mar. 2022). URL: https://hal.inria.fr/hal-02006471.

[12] C. Brécheteau, E. Genetay, T. Mathieu and A. Saumard. 'Topics in robust statistical learning'. In: *ESAIM: Proceedings and Surveys* (2022). URL: https://hal.archives-ouvertes.fr/hal-03605702.

[13] R. Gautron, O.-A. Maillard, P. Preux, M. Corbeels and R. Sabbadin. 'Reinforcement Learning for crop management'. In: *Computers and Electronics in Agriculture* 200 (27th July 2022), p. 107182. DOI: 10.1016/j.compag.2022.107182. URL: https://hal.inria.fr/hal-03834290.

[14] T. Mathieu. 'Concentration study of M-estimators using the influence function'. In: *Electronic Journal of Statistics* 16.1 (1st Jan. 2022), pp. 3695–3750. DOI: 10.1214/22-ejs2030. URL: https://hal.archives-ouvertes.fr/hal-03757720.

[15] E. Ménager, P. Schegg, E. Khairallah, D. Marchal, J. Dequidt, P. Preux and C. Duriez. 'SofaGym: An open platform for Reinforcement Learning based on Soft Robot simulations'. In: *Soft Robotics* (2022). URL: https://hal.inria.fr/hal-03778189.

International peer-reviewed conferences

[16] A. Azize and D. Basu. 'When Privacy Meets Partial Information: A Refined Analysis of Differentially Private Bandits'. In: Advances in Neural Information Processing Systems. New Orleans, United States, Dec. 2022. URL: https://hal.archives-ouvertes.fr/hal-03781600.

[17] D. Baudry, Y. Russac and E. Kaufmann. 'Efficient Algorithms for Extreme Bandits'. In: International conferenece on Articifial Intelligence and Statistics (AISTATS). Proceedings of Machine Learning Research (PMLR). Virtual Conference, Spain, 28th Mar. 2022. URL: https://hal.archives-ouvertes.fr/hal-03741302.

[18] O. Beaumont, L. Eyraud-Dubois and A. Shilova. 'MadPipe: Memory Aware Dynamic Programming Algorithm for Pipelined Model Parallelism'. In: ScaDL 2022 - Scalable Deep Learning over Parallel and Distributed Infrastructure - An IPDPS 2022 Workshop. Proceedings of IPDPS W'22. Lyon / Virtual, France, 2022. URL: https://hal.archives-ouvertes.fr/hal-03025305.

[19] T. K. Buening, M. Segal, D. Basu, A.-M. George and C. Dimitrakakis. 'On Meritocracy in Optimal Set Selection'. In: EAAMO 2022- Equity and Access in Algorithms, Mechanisms, and Optimization. Arlington, United States, 17th Oct. 2022. URL: https://hal.archives-ouvertes.fr/hal-03445971.

[20] H. Eriksson, D. Basu, M. Alibeigi and C. Dimitrakakis. 'SENTINEL: Taming Uncertainty with Ensemble-based Distributional Reinforcement Learning'. In: UAI 2022- Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence. Vol. 180. Proceedings of Machine Learning Research. Eindhoven, Netherlands, Aug. 2022, pp. 631–640. URL: https://hal.archives-ouvertes.fr/hal-03150823.

[21] Y. Flet-Berliac and D. Basu. 'SAAC: Safe Reinforcement Learning as an Adversarial Game of Actor-Critics'. In: RLDM 2022 - The Multi-disciplinary Conference on Reinforcement Learning and Decision Making. Providence, United States, 8th June 2022. URL: https://hal.archives-ouvertes.fr/hal-03771734.

[22] B. Ghosh, D. Basu and K. S. Meel. 'Algorithmic fairness verification with graphical models'. In: AAAI-2022 - 36th AAAI Conference on Artificial Intelligence. Vol. 2. Virtual, United States, 2022. URL: https://hal.archives-ouvertes.fr/hal-03770361.

[23] M. Jourdan and R. Degenne. 'Choosing Answers in epsilon-Best-Answer Identification for Linear Bandits'. In: 39th International Conference on Machine Learning (ICML 2022). Baltimore, United States, 17th July 2022. URL: https://hal.inria.fr/hal-03830700.

[24] 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.archives-ouvertes.fr/hal-03825103.

[25] M. H. Nguyen, L. Sun-Hosoya, N. Grinsztajn and I. Guyon. 'Meta-learning from Learning Curves: Challenge Design and Baseline Results'. In: IJCNN 2022 - International Joint Conference on Neural Networks. Padua, Italy: IEEE, 18th July 2022, pp. 1–8. URL: https://hal.archives-ouvertes.fr/hal-03740118.

[26] 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.archives-ouvertes.fr/hal-03825423.

[27] C. Réda, S. Vakili and E. Kaufmann. 'Near-Optimal Collaborative Learning in Bandits'. In: NeurIPS 2022 - 36th Conference on Neural Information Processing System. Advances in Neural Processing Systems. New Orleans, United States, Dec. 2022. URL: https://hal.archives-ouvertes.fr/hal-03825099.

[28] S. Rezaeifar, R. Dadashi, N. Vieillard, L. Hussenot, O. Bachem, O. Pietquin and M. Geist. 'Offline Reinforcement Learning as Anti-Exploration'. In: AAAI 2022 - 36th AAAI Conference on Artificial Intelligence. Vancouver, Canada, 22nd Feb. 2022. URL: https://hal.archives-ouvertes.fr/hal-03468875.

[29] P. Schegg, J. Dequidt, E. Coevoet, E. Leurent, R. Sabatier, P. Preux and C. Duriez. 'Automated planning for robotic guidewire navigation in the coronary arteries'. In: Robosoft 2022 - International Conference on Soft Robotics. Edimbourg, United Kingdom, 13th Apr. 2022. URL: https://hal.inria.fr/hal-03778352.

[30] A. Tirinzoni and R. Degenne. 'On Elimination Strategies for Bandit Fixed-Confidence Identification'. In: NeurIPS 2022 - 36th Conference on Neural Information Processing System. New Orleans, United States, 28th Nov. 2022. URL: https://hal.inria.fr/hal-03830692.

[31] A. Tirinzoni, A. Al-Marjani and E. Kaufmann. 'Near Instance-Optimal PAC Reinforcement Learning for Deterministic MDPs'. In: NeurIPS 2022 - 36th Conference on Neural Information Processing System. Vol. Advances in Neural Information Processing Systems. New Orleans, United States, 28th Nov. 2022. URL: https://hal.archives-ouvertes.fr/hal-03825101.

[32] J. Wang, D. Basu and I. Trummer. 'Procrastinated Tree Search: Black-box Optimization with Delayed, Noisy, and Multi-fidelity Feedback'. In: AAAI Conference on Artificial Intelligence. Vol. 36. Proceedings of the AAAI Conference on Artificial Intelligence 9. Virtual, United States, 28th June 2022, pp. 10381–10390. URL: https://hal.archives-ouvertes.fr/hal-03445909.

[33] J. Wang, I. Trummer and D. Basu. 'UDO: Universal Database Optimization using Reinforcement Learning'. In: Proceedings of the VLDB Endowment. Vol. 14. Proceedings of the VLDB Endowment 13. Sydney, Australia: VLDB Endowment, Sept. 2021, pp. 3402–3414. DOI: 10.14778/3484224.3484236. URL: https://hal.archives-ouvertes.fr/hal-03445686.

Conferences without proceedings

[34] H. Eriksson, D. Basu, M. Alibeigi and C. Dimitrakakis. 'Risk-Sensitive Bayesian Games for Multi-Agent Reinforcement Learning under Policy Uncertainty'. In: OptLearnMAS@AAMAS. Workshop on Optimization and Learning in Multiagent Systems at International Conference on Autonomous Agents and Multiagent Systems. Virtual, New Zealand, May 2022. URL: https://hal.archives-ouvertes.fr/hal-03770369.

[35] R. Ouhamma, D. Basu and O.-A. Maillard. 'Bilinear Exponential Family of MDPs: Frequentist Regret Bound with Tractable Exploration & Planning'. In: EWRL 2022 – European Workshop on Reinforcement Learning. Milan, Italy, Sept. 2022. URL: https://hal.archives-ouvertes.fr/hal-03790997.

[36] P. Saux and O.-A. Maillard. 'Risk-aware linear bandits with convex loss'. In: European Workshop on Reinforcement Learning. Milan, Italy, 19th Sept. 2022. URL: https://hal.archives-ouvertes.fr/hal-03776680.

[37] A. Tirinzoni, A. Al-Marjani and E. Kaufmann. 'Optimistic PAC Reinforcement Learning: the Instance-Dependent View'. In: EWRL 2022 - European Workshop on Reinforcement Learning. Milan, Italy, 19th Sept. 2022. URL: https://hal.archives-ouvertes.fr/hal-03767409.

Scientific book chapters

[38] N. Mitton, L. Brossard, T. Bouadi, F. Garcia, R. Gautron, N. Hilgert, D. Ienco, C. Largouët, E. Lutton, V. Masson, R. Martin-Clouaire, M.-L. Mugnier, P. Neveu, P. Preux, H. Raynal, C. Roussey, A. Termier and V. Bellon Maurel. 'Foundations and state of play'. In: *Agriculture and Digital Technology: Getting the most out of digital technology to contribute to the transition to sustainable agriculture and food systems*. White book Inrira 6. INRIA, 2022, pp. 30–75. URL: https://hal.inrae.fr/hal-03609470.

Doctoral dissertations and habilitation theses

[39] O. D. Domingues. 'Exploration in Reinforcement Learning: Beyond Finite State-Spaces'. Université de Lille, 18th Mar. 2022. URL: https://theses.hal.science/tel-03720236.

[40] C. Réda. 'Combination of gene regulatory networks and sequential machine learning for drug repurposing'. Université Paris Cité, 9th Sept. 2022. URL: https://hal.archives-ouvertes.fr/tel-03846072.

Reports & preprints

[41] D. Basu, O.-A. Maillard and T. Mathieu. *Bandits Corrupted by Nature: Lower Bounds on Regret and Robust Optimistic Algorithm*. 17th Mar. 2022. URL: https://hal.archives-ouvertes.fr/hal-03611816.

[42] O. Beaumont, L. Eyraud-Dubois and A. Shilova. *An Integer Linear Programming Approach for Pipelined Model Parallelism*. RR-9452. Inria, Jan. 2022. URL: https://hal.inria.fr/hal-03549009.

[43] O. Beaumont, L. Eyraud-Dubois, A. Shilova and X. Zhao. *Weight Offloading Strategies for Training Large DNN Models*. 18th Feb. 2022. URL: https://hal.inria.fr/hal-03580767.

[44] R. Della Vecchia and D. Basu. *Online Instrumental Variable Regression: Regret Analysis and Bandit Feedback*. 26th Oct. 2022. URL: https://hal.archives-ouvertes.fr/hal-03831210.

[45] R. Della Vecchia, A. Shilova, P. Preux and R. Akrour. *Entropy Regularized Reinforcement Learning with Cascading Networks*. 7003. Inria Lille Nord Europe - Laboratoire CRIStAL - Université de Lille, 19th Sept. 2022, p. 16. URL: https://hal.archives-ouvertes.fr/hal-03793130.

[46] R. Gautron, E. J. Padrón, P. Preux, J. Bigot, O.-A. Maillard and D. Emukpere. *gym-DSSAT: a crop model turned into a Reinforcement Learning environment*. RR-9460. Inria Lille, 1st July 2022, p. 31. URL: https://hal.inria.fr/hal-03711132.

[47] B. Ghosh, D. Basu and K. S. Meel. *How Biased is Your Feature?: Computing Fairness Influence Functions with Global Sensitivity Analysis*. 6th Sept. 2022. URL: https://hal.archives-ouvertes.fr/hal-03770346.

[48] M. Jourdan and R. Degenne. *Non-Asymptotic Analysis of a UCB-based Top Two Algorithm*. 26th Oct. 2022. URL: https://hal.inria.fr/hal-03830958.

[49] M. H. Nguyen, L. Sun, N. Grinsztajn and I. Guyon. *Meta-learning from Learning Curves Challenge: Lessons learned from the First Round and Design of the Second Round*. 3rd Aug. 2022. URL: https://hal.archives-ouvertes.fr/hal-03725313.

12.3 Cited publications

[50] M. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, 1994.

[51] B. Recht. 'A Tour of Reinforcement Learning: The View from Continuous Control'. arxiv preprint 1806.09460. 2018.

[52] R. Sutton and A. Barto. *Reinforcement Learning: an Introduction*. 2nd ed. http://incompleteideas.net/book/the-book-2nd.html. MIT Press, 2018.

[53] C. Szepesvári and T. Lattimore. *Bandit Algorithms*. Cambridge University press, 2019.