RESEARCH CENTRE

Lille - Nord Europe

IN PARTNERSHIP WITH:

CNRS, Université de Lille

2021 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

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

- Debabrota Basu [Inria, Starting Faculty Position]
- Remy Degenne [Inria, Starting Faculty Position]
- Emilie Kaufmann [CNRS, Researcher, HDR]
- Odalric Maillard [Inria, Researcher, HDR]
- Jill Jenn Vie [Inria, Researcher, left Scool at the end of Oct. 2021]

Faculty Member

- Philippe Preux [Team leader, Université de Lille, Professor, HDR]

Post-Doctoral Fellows

- Rianne De Heide [Inria, in Scool since Sep 2021]
- Riccardo Della Vecchia [Inria, in Scool since Sep 2021]
- Timothee Mathieu [Inria, in Scool since Oct 2021]
- Sein Minn [Inria]
- Mohit Mittal [Inria]
- Andrea Tirinzoni [Inria, in Scool since Apr 2021]

PhD Students

- Achraf Azize [Université de Lille, in Scool since Oct 2021]
- Dorian Baudry [CNRS]
- Omar Darwiche Domingues [Inria]
- Johan Ferret [Google]
- Yannis Flet-Berliac [Université de Lille, left Scool in Sep 2021]
- Nathan Grinsztajn [École polytechnique]
- Leonard Hussenot-Desenonges [Google]
- Marc Jourdan [Université de Lille, in Scool as an intern since Apr 2021, then PhD student since Oct 2021]
- Matheus Medeiros Centa [Université de Lille, in Scool since Apr 2021 as an intern, then as a PhD student]
- Reda Ouhamma [École polytechnique]
- Sarah Perrin [Université de Lille]
- Fabien Pesquerel [École Normale Supérieure de Paris]
- Clemence Reda [Université de Paris]
- Hassan Saber [Inria, from Sep 2021]

- Patrick Saux [Inria]
- Mathieu Seurin [Université de Lille, left Scool at the end of Aug 2021]
- Xuedong Shang [Université de Lille, left Scool at the end of June 2021]
- Jean Tarbouriech [Facebook]

Technical Staff

- Hernan David Carvajal Bastidas [Inria, Engineer, in Scool since Mar 2021]
- David Emukpere [Inria, Engineer, in Scool since Nov 2021]
- Clemence Leguillette [Inria, Engineer, left Scool at the end of Jan. 2021]
- Antoine Moulin-Bonno [Inria, Engineer, in Scool from Feb 2021 until Apr 2021]
- Vianney Taquet [Inria, Engineer, left Scool at the end of Jan. 2021]
- Julien Teigny [Inria, Engineer]

Interns and Apprentices

- Thomas Carta [Inria, in Scool from July to Sep.]
- Florent Dufay [École normale supérieure de Rennes, in Scool in June and Jul 2021]
- Toby Johnstone [Université de Lille, in Sccol from Apr. to Sep.]
- Marc Jourdan [Inria, from Mar 2021 until Jul 2021]
- Danilo Moreira Lagos [Inria, in Scool from Sep. to Dec.]

Administrative Assistant

- Amelie Supervielle [Inria]

Visiting Scientist

- Bishwamittra Ghosh [Université nationale de Singapour, in Scool since Nov 2021]

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 will consider bandit problems and the reinforcement learning (RL) problem. In a simplified way, RL considers the problem of learning an optimal policy in a Markov Decision Problem (MDP); 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. We also consider online learning where the goal is to learn a model from a stream of data, such as learning a compressed representation of a stream of data (each data may be a scalar, a vector, or even a more complex data structure such as a tree or a graph). 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.

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 2021, 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 common 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 ad 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

- É. Leurent was awarded the Ph.D. prize by CNRS (GdR MACS) and club EEA, that recognizes the best research work on Modelling, Analysis and Control of Dynamical Systems for his dissertation defended in October 2020. Award tied with Eva Petitdemange.
- É. Leurent was awarded the Ph.D. prize by Sanef-Abertis Chair, that recognizes the best research work on Road Safety for his dissertation defended in October 2020.

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: Omar Darwiche Domingues

7.1.2 justicia

Name: Justicia: A Stochastic SAT Approach to Formally Verify Fairness

Keywords: Fairness, Machine learning, Verification, Fairness Verification, Fair and ethical machine learning, Formal methods

Functional Description: justicia is a fairness verifier written in Python. The library provides a stochastic SAT encoding of multiple fairness definitions and fair ML algorithms. justicia then further verifies the fairness metric achieved by the corresponding ML algorithm. It is now available as an official Python package and can be installed using pip.

News of the Year: 2020

URL: https://www.github.com/meelgroup/justicia

Contact: Debabrota Basu

Participant: Bishwamittra Ghosh

Partner: National University of Singapore

7.1.3 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: http://gitlab.inria.fr/scool/gym-dssat

Contact: Romain Gautron

Partners: CIRAD, Cgiar

8 New results

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

8.1 Bandit problems

Sample complexity bounds for stochastic shortest path with a generative model, [50]

We consider the objective of computing an ϵ-optimal policy in a stochastic shortest path (SSP) setting, provided that we can access a generative sampling oracle. We propose two algorithms for this setting and derive PAC bounds on their sample complexity: one for the case of positive costs and the other for the case of non-negative costs under a restricted optimality criterion. While tight sample complexity bounds have been derived for the finite-horizon and discounted MDPs, the SSP problem is a strict generalization of these settings and it poses additional technical challenges due to the fact that no specific time horizon is prescribed and policies may never terminate, i.e., we are possibly facing non-proper policies. As a consequence, we can neither directly apply existing techniques minimizing sample complexity nor rely on a regret-to-PAC conversion leveraging recent regret bounds for SSP. Our analysis instead combines SSP-specific tools and variance reduction techniques to obtain the first sample complexity bounds for this setting.

Routine Bandits: Minimizing Regret on Recurring Problems, [47]

We study a variant of the multi-armed bandit problem in which a learner faces every day one of B many bandit instances, and call it a routine bandit. More specifically, at each period h ∈ [1, H] , the same bandit b^h is considered during T > 1 consecutive time steps, but the identity b^h is unknown to the learner. We assume all rewards distribution are Gaussian standard. Such a situation typically occurs in recommender systems when a learner may repeatedly serve the same user whose identity is unknown due to privacy issues. By combining banditidentification tests with a KLUCB type strategy, we introduce the KLUCB for Routine Bandits (KLUCB-RB) algorithm. While independently running KLUCB algorithm at each period leads to a cumulative expected regret of $\Omega(H \log T)$ after H many periods when $T \to \infty$, KLUCB-RB benefits from previous periods by aggregating observations from similar identified bandits, which yields a non-trivial scaling of $\Omega(\log T)$. This is achieved without knowing which bandit instance is being faced by KLUCB-RB on this period, nor knowing a priori the number of possible bandit instances. We provide numerical illustration that confirm the benefit of KLUCB-RB while using less information about the problem compared with existing strategies for similar problems.

Non-Asymptotic Sequential Tests for Overlapping Hypotheses and application to near optimal arm identification in bandit models, [12]

In this paper, we study sequential testing problems with overlapping hypotheses. We first focus on the simple problem of assessing if the mean μ of a Gaussian distribution is ≥ ϵ– or ≤ ϵ; if $\mu \in (--\epsilon, \epsilon)$, both answers are considered to be correct. Then, we consider PAC-best arm identification in a bandit model: given K probability distributions on R with means $\mu_1, ..., \mu_K$, we derive the asymptotic complexity of identifying, with risk at most δ, an index $I \in 1, ..., K$ such that $\mu_I \geq max_i \mu_i - -\epsilon$. We provide non asymptotic bounds on the error of a parallel General Likelihood Ratio Test, which can also be used for more general testing problems. We further propose lower bound on the number of observation needed to identify a correct hypothesis. Those lower bounds rely on information-theoretic arguments, and specifically on two versions of a change of measure lemma (a high-level form, and a low-level form) whose relative merits are discussed.

From Optimality to Robustness: Dirichlet Sampling Strategies in Stochastic Bandits, [19]

The stochastic multi-arm bandit problem has been extensively studied under standard assumptions on the arm's distribution (e.g bounded with known support, exponential family, etc). These assumptions are suitable for many real-world problems but sometimes they require knowledge (on tails for instance) that may not be precisely accessible to the practitioner, raising the question of the robustness of bandit algorithms to model misspecification. In this paper we study a generic Dirichlet Sampling (DS) algorithm, based on pairwise comparisons of empirical indices computed with re-sampling of the arms' observations and a data-dependent exploration bonus. We show that different variants of this strategy achieve provably optimal regret guarantees when the distributions are bounded and logarithmic regret for semi-bounded distributions with a mild quantile condition. We also show that a simple tuning achieve robustness with respect to a large class of unbounded distributions, at the cost of slightly worse than logarithmic asymptotic regret. We finally provide numerical experiments showing the merits of DS in a decision-making problem on synthetic agriculture data.

Top-m identification for linear bandits, [43]

Motivated by an application to drug repurposing, we propose the first algorithms to tackle the

identification of the m ≥ 1 arms with largest means in a linear bandit model, in the fixed-confidence setting. These algorithms belong to the generic family of Gap-Index Focused Algorithms (GIFA) that we introduce for Top-m identification in linear bandits. We propose a unified analysis of these algorithms, which shows how the use of features might decrease the sample complexity. We further validate these algorithms empirically on simulated data and on a simple drug repurposing task.

Dealing With Misspecification In Fixed-Confidence Linear Top-m Identification, [44]

We study the problem of the identification of m arms with largest means under a fixed error rate δ (fixed-confidence Top-m identification), for misspecified linear bandit models. This problem is motivated by practical applications, especially in medicine and recommendation systems, where linear models are popular due to their simplicity and the existence of efficient algorithms, but in which data inevitably deviates from linearity. In this work, we first derive a tractable lower bound on the sample complexity of any δ-correct algorithm for the general Top-m identification problem. We show that knowing the scale of the deviation from linearity is necessary to exploit the structure of the problem. We then describe the first algorithm for this setting, which is both practical and adapts to the amount of misspecification. We derive an upper bound to its sample complexity which confirms this adaptivity and that matches the lower bound when $\delta \to 0$. Finally, we evaluate our algorithm on both synthetic and real-world data, showing competitive performance with respect to existing baselines.

On Limited-Memory Subsampling Strategies for Bandits, [18]

There has been a recent surge of interest in nonparametric bandit algorithms based on subsampling. One drawback however of these approaches is the additional complexity required by random subsampling and the storage of the full history of rewards. Our first contribution is to show that a simple deterministic subsampling rule, proposed in the recent work of Baudry et al. (2020) under the name of "last-block subsampling", is asymptotically optimal in one-parameter exponential families. In addition, we prove that these guarantees also hold when limiting the algorithm memory to a polylogarithmic function of the time horizon. These findings open up new perspectives, in particular for non-stationary scenarios in which the arm distributions evolve over time. We propose a variant of the algorithm in which only the most recent observations are used for subsampling, achieving optimal regret guarantees under the assumption of a known number of abrupt changes. Extensive numerical simulations highlight the merits of this approach, particularly when the changes are not only affecting the means of the rewards. **Optimal Thompson Sampling strategies for support-aware CVaR bandits**, [17]

In this paper we study a multi-arm bandit problem in which the quality of each arm is measured by the Conditional Value at Risk (CVaR) at some level alpha of the reward distribution. While existing works in this setting mainly focus on Upper Confidence Bound algorithms, we introduce a new Thompson Sampling approach for CVaR bandits on bounded rewards that is flexible enough to solve a variety of problems grounded on physical resources. Building on a recent work by Riou & Honda (2020), we introduce B-CVTS for continuous bounded rewards and M-CVTS for multinomial distributions. On the theoretical side, we provide a non-trivial extension of their analysis that enables to theoretically bound their CVaR regret minimization performance. Strikingly, our results show that these strategies are the first to provably achieve asymptotic optimality in CVaR bandits, matching the corresponding asymptotic lower bounds for this setting. Further, we illustrate empirically the benefit of Thompson Sampling approaches both in a realistic environment simulating a use-case in agriculture and on various synthetic examples.

Online Sign Identification: Minimization of the Number of Errors in Thresholding Bandits, [39]

In the fixed budget thresholding bandit problem, an algorithm sequentially allocates a budgeted number of samples to different distributions. It then predicts whether the mean of each distribution is larger or lower than a given threshold. We introduce a large family of algorithms (containing most existing relevant ones), inspired by the Frank-Wolfe algorithm, and provide a thorough yet generic analysis of their performance. This allowed us to construct new explicit algorithms, for a broad class of problems, whose losses are within a small constant factor of the non-adaptive oracle ones. Quite interestingly, we observed that adaptive methods empirically greatly out-perform non-adaptive oracles, an uncommon behavior in standard online learning settings, such as regret minimization. We explain this surprising phenomenon on an insightful toy problem.

Stochastic bandits with groups of similar arms, [42]

We consider a variant of the stochastic multi-armed bandit problem where arms are known to be organized into different groups having the same mean. The groups are unknown but a lower bound q on their size is known. This situation typically appears when each arm can be described with a list of categorical

attributes, and the (unknown) mean reward function only depends on a subset of them, the others being redundant. In this case, q is linked naturally to the number of attributes considered redundant, and the number of categories of each attribute. For this structured problem of practical relevance, we first derive the asymptotic regret lower bound and corresponding constrained optimization problem. They reveal the achievable regret can be substantially reduced when compared to the unstructured setup, possibly by a factor q. However, solving exactly the exact constrained optimization problem involves a combinatorial problem. We introduce a lowerbound inspired strategy involving a computationally efficient relaxation that is based on a sorting mechanism. We further prove it achieves a lower bound close to the optimal one up to a controlled factor, and achieves an asymptotic regret q times smaller than the unstructured one. We believe this shows it is a valuable strategy for the practitioner. Last, we illustrate the performance of the considered strategy on numerical experiments involving a large number of arms.

Indexed Minimum Empirical Divergence for Unimodal Bandits, [46]

We consider a multi-armed bandit problem specified by a set of one-dimensional family exponential distributions endowed with a unimodal structure. We introduce IMED-UB, a algorithm that optimally exploits the unimodal-structure, by adapting to this setting the Indexed Minimum Empirical Divergence (IMED) algorithm introduced by Honda and Takemura [2015]. Owing to our proof technique, we are able to provide a concise finite-time analysis of IMED-UB algorithm. Numerical experiments show that IMED-UB competes with the state-of-the-art algorithms.

Procrastinated Tree Search: Black-box Optimization with Delayed, Noisy, and Multi-fidelity Feedback, [62]

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 $O(\log T)$ and worsens by $O(T^{\frac{1-\alpha}{d+2}})$ 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.

8.2 Reinforcement learning

What Matters In On-Policy Reinforcement Learning? A Large-Scale Empirical Study, [16]

In recent years, on-policy reinforcement learning (RL) has been successfully applied to many different continuous control tasks. While RL algorithms are often conceptually simple, their state-of-the-art implementations take numerous low- and high-level design decisions that strongly affect the performance of the resulting agents. Those choices are usually not extensively discussed in the literature, leading to discrepancy between published descriptions of algorithms and their implementations. This makes it hard to attribute progress in RL and slows down overall progress [Engstrom'20]. As a step towards filling that gap, we implement >50 such "choices" in a unified on-policy RL framework, allowing us to investigate their impact in a large-scale empirical study. We train over 250'000 agents in five continuous control environments of different complexity and provide insights and practical recommendations for on-policy training of RL agents.

Show me the Way: Intrinsic Motivation from Demonstrations, [34]

The study of exploration in the domain of decision making has a long history but remains actively debated. From the vast literature that addressed this topic for decades under various points of view (e.g., developmental psychology, experimental design, artificial intelligence), intrinsic motivation emerged as a concept that can practically be transferred to artificial agents. Especially, in the recent field of Deep Reinforcement Learning (RL), agents implement such a concept (mainly using a novelty argument) in the shape of an exploration bonus, added to the task reward, that encourages visiting the whole environment.

This approach is supported by the large amount of theory on RL for which convergence to optimality assumes exhaustive exploration. Yet, Human Beings and mammals do not exhaustively explore the world and their motivation is not only based on novelty but also on various other factors (e.g., curiosity, fun, style, pleasure, safety, competition, etc.). They optimize for life-long learning and train to learn transferable skills in playgrounds without obvious goals. They also apply innate or learned priors to save time and stay safe. For these reasons, we propose to learn an exploration bonus from demonstrations that could transfer these motivations to an artificial agent with little assumptions about their rationale. Using an inverse RL approach, we show that complex exploration behaviors, reflecting different motivations, can be learnt and efficiently used by RL agents to solve tasks for which exhaustive exploration is prohibitive.

Primal Wasserstein Imitation Learning, [22]

Imitation Learning (IL) methods seek to match the behavior of an agent with that of an expert. In the present work, we propose a new IL method based on a conceptually simple algorithm: Primal Wasserstein Imitation Learning (PWIL), which ties to the primal form of the Wasserstein distance between the expert and the agent state-action distributions. We present a reward function which is derived offline, as opposed to recent adversarial IL algorithms that learn a reward function through interactions with the environment, and which requires little fine-tuning. We show that we can recover expert behavior on a variety of continuous control tasks of the MuJoCo domain in a sample efficient manner in terms of agent interactions and of expert interactions with the environment. Finally, we show that the behavior of the agent we train matches the behavior of the expert with the Wasserstein distance, rather than the commonly used proxy of performance.

Self-Imitation Advantage Learning, [27]

Self-imitation learning is a Reinforcement Learning (RL) method that encourages actions whose returns were higher than expected, which helps in hard exploration and sparse reward problems. It was shown to improve the performance of on-policy actor-critic methods in several discrete control tasks. Nevertheless, applying self-imitation to the mostly action-value based off-policy RL methods is not straightforward. We propose SAIL, a novel generalization of self-imitation learning for off-policy RL, based on a modification of the Bellman optimality operator that we connect to Advantage Learning. Crucially, our method mitigates the problem of stale returns by choosing the most optimistic return estimate between the observed return and the current action-value for self-imitation. We demonstrate the empirical effectiveness of SAIL on the Arcade Learning Environment, with a focus on hard exploration games.

Kernel-based reinforcement Learning: A finite-time analysis, [25]

We consider the exploration-exploitation dilemma in finite-horizon reinforcement learning problems whose state-action space is endowed with a metric. We introduce Kernel-UCBVI, a model-based optimistic algorithm that leverages the smoothness of the MDP and a non-parametric kernel estimator of the rewards and transitions to efficiently balance exploration and exploitation. Unlike existing approaches with regret guarantees, it does not use any kind of partitioning of the state-action space. For problems with K episodes and horizon H, we provide a regret bound of $O(H^3 K^{-\frac{(2d}{2d+1)}})$, where d is the covering dimension of the joint state-action space. This is the first regret bound for kernel-based RL using smoothing kernels, which requires very weak assumptions on the MDP and has been previously applied to a wide range of tasks. We empirically validate our approach in continuous MDPs with sparse rewards.

A kernel-based approach to non-stationary reinforcement learning in metric spaces, [24]

In this work, we propose KeRNS: an algorithm for episodic reinforcement learning in nonstationary Markov Decision Processes (MDPs) whose state-action set is endowed with a metric. Using a non-parametric model of the MDP built with time-dependent kernels, we prove a regret bound that scales with the covering dimension of the state-action space and the total variation of the MDP with time, which quantifies its level of non-stationarity. Our method generalizes previous approaches based on sliding windows and exponential discounting used to handle changing environments. We further propose a practical implementation of KeRNS, we analyze its regret and validate it experimentally.

Don't Do What Doesn't Matter: Intrinsic Motivation with Action Usefulness, [48]

Sparse rewards are double-edged training signals in reinforcement learning: easy to design but hard to optimize. Intrinsic motivation guidances have thus been developed toward alleviating the resulting exploration problem. They usually incentivize agents to look for new states through novelty signals. Yet, such methods encourage exhaustive exploration of the state space rather than focusing on the

environment's salient interaction opportunities. We propose a new exploration method, called Don't Do What Doesn't Matter (DoWhaM), shifting the emphasis from state novelty to state with relevant actions. While most actions consistently change the state when used, e.g. moving the agent, some actions are only effective in specific states, e.g., opening a door, grabbing an object. DoWhaM detects and rewards actions that seldom affect the environment. We evaluate DoWhaM on the procedurallygenerated environment MiniGrid, against state-ofthe-art methods. Experiments consistently show that DoWhaM greatly reduces sample complexity, installing the new state-of-the-art in MiniGrid.

Learning Value Functions in Deep Policy Gradients using Residual Variance, [29]

Policy gradient algorithms have proven to be successful in diverse decision making and control tasks. However, these methods suffer from high sample complexity and instability issues. In this paper, we address these challenges by providing a different approach for training the critic in the actor-critic framework. Our work builds on recent studies indicating that traditional actor-critic algorithms do not succeed in fitting the true value function, calling for the need to identify a better objective for the critic. In our method, the critic uses a new state-value (resp. state-action-value) function approximation that learns the value of the states (resp. state-action pairs) relative to their mean value rather than the absolute value as in conventional actor-critic. We prove the theoretical consistency of the new gradient estimator and observe dramatic empirical improvement across a variety of continuous control tasks and algorithms. Furthermore, we validate our method in tasks with sparse rewards, where we provide experimental evidence and theoretical insights.

READYS: A Reinforcement Learning Based Strategy for Heterogeneous Dynamic Scheduling, [31]

In this paper, we propose READYS, a reinforcement learning algorithm for the dynamic scheduling of computations modeled as a Directed Acyclic Graph (DAGs). Our goal is to develop a scheduling algorithm in which allocation and scheduling decisions are made at runtime, based on the state of the system, as performed in runtime systems such as StarPU or ParSEC. Reinforcement Learning is a natural candidate to achieve this task, since its general principle is to build step by step a strategy that, given the state of the system (the state of the resources and a view of the ready tasks and their successors in our case), makes a decision to optimize a global criterion. Moreover, the use of Reinforcement Learning is natural in a context where the duration of tasks (and communications) is stochastic. We propose READYS that combines Graph Convolutional Networks (GCN) with an Actor-Critic Algorithm (A2C): it builds an adaptive representation of the scheduling problem on the fly and learns a scheduling strategy, aiming at minimizing the makespan. A crucial point is that READYS builds a general scheduling strategy which is neither limited to only one specific application or task graph nor one particular problem size, and that can be used to schedule any DAG. We focus on different types of task graphs originating from linear algebra factorization kernels (CHOLESKY, LU, QR) and we consider heterogeneous platforms made of a few CPUs and GPUs. We first propose to analyze the performance of READYS when learning is performed on a given (platform, kernel, problem size) combination. Using simulations, we show that the scheduling agent obtains performances very similar or even superior to algorithms from the literature, and that it is especially powerful when the scheduling environment contains a lot of uncertainty. We additionally demonstrate that our agent exhibits very promising generalization capabilities. To the best of our knowledge, this is the first paper which shows that reinforcement learning can really be used for dynamic DAG scheduling on heterogeneous resources.

There Is No Turning Back: A Self-Supervised Approach for Reversibility-Aware Reinforcement Learning, [32]

We propose to learn to distinguish reversible from irreversible actions for better informed decision-making in Reinforcement Learning (RL). From theoretical considerations, we show that approximate reversibility can be learned through a simple surrogate task: ranking randomly sampled trajectory events in chronological order. Intuitively, pairs of events that are always observed in the same order are likely to be separated by an irreversible sequence of actions. Conveniently, learning the temporal order of events can be done in a fully self-supervised way, which we use to estimate the reversibility of actions from experience, without any priors. We propose two different strategies that incorporate reversibility in RL agents, one strategy for exploration (RAE) and one strategy for control (RAC). We demonstrate the potential of reversibility-aware agents in several environments, including the challenging Sokoban game. In synthetic tasks, we show that we can learn control policies that never fail and reduce to zero the side-effects of interactions, even without access to the reward function.

Adaptive reward-free exploration, [35]

Reward-free exploration is a reinforcement learning setting recently studied by Jin et al., who address it by running several algorithms with regret guarantees in parallel. In our work, we instead propose a more adaptive approach for reward-free exploration which directly reduces upper bounds on the maximum MDP estimation error. We show that, interestingly, our reward-free UCRL algorithm can be seen as a variant of an algorithm of Fiechter from 1994 [11], originally proposed for a different objective that we call best-policy identification. We prove that RF-UCRL needs O (SAH 4 /ϵ 2) log(1/δ) episodes to output, with probability 1 – δ, an ϵ-approximation of the optimal policy for any reward function. We empirically compare it to oracle strategies using a generative model.

Fast active learning for pure exploration in reinforcement learning, [36]

Realistic environments often provide agents with very limited feedback. When the environment is initially unknown, the feedback, in the beginning, can be completely absent, and the agents may first choose to devote all their effort on exploring efficiently. The exploration remains a challenge while it has been addressed with many hand-tuned heuristics with different levels of generality on one side, and a few theoretically backed exploration strategies on the other. Many of them are incarnated by *intrinsic motivation* and in particular *explorations bonuses*. A common rule of thumb for exploration bonuses is to use $1/\sqrt{n}$ bonus that is added to the empirical estimates of the reward, where n is a number of times this particular state (or a state-action pair) was visited. We show that, surprisingly, for a pure-exploration objective of reward-free exploration, bonuses that scale with 1/n bring faster learning rates, improving the known upper bounds with respect to the dependence on the horizon H. Furthermore, we show that with an improved analysis of the stopping time, we can improve by a factor H the sample complexity in the *best-policy identification* setting, which is another pure-exploration objective, where the environment provides rewards but the agent is not penalized for its behavior during the exploration phase.

UCB Momentum Q-learning: Correcting the bias without forgetting, [37]

We propose UCBMQ, Upper Confidence Bound Momentum Q-learning, a new algorithm for reinforcement learning in tabular and possibly stagedependent, episodic Markov decision process. UCBMQ is based on Q-learning where we add a momentum term and rely on the principle of optimism in face of uncertainty to deal with exploration. Our new technical ingredient of UCBMQ is the use of momentum to correct the bias that Q-learning suffers while, at the same time, limiting the impact it has on the the second-order term of the regret. For UCBMQ, we are able to guarantee a regret of at most $O(\sqrt{H3SAT + H4SA})$ where H is the length of an episode, S the number of states, A the number of actions, T the number of episodes and ignoring terms in poly log(SAHT). Notably, UCBMQ is the first algorithm that simultaneously matches the lower bound of $\Omega(\sqrt{H3SAT})$ for large enough T and has a second-order term (with respect to the horizon T) that scales only linearly with the number of states S.

Episodic reinforcement learning in finite MDPs: Minimax lower bounds revisited, [26]

In this paper, we propose new problem-independent lower bounds on the sample complexity and regret in episodic MDPs, with a particular focus on the non-stationary case in which the transition kernel is allowed to change in each stage of the episode. Our main contribution is a lower bound of $\Omega((H3SA/\epsilon 2)log(1/\delta))$ on the sample complexity of an (ϵ,δ)-PAC algorithm for best policy identification in a non-stationary MDP, relying on a construction of "hard MDPs" which is different from the ones previously used in the literature. Using this same class of MDPs, we also provide a rigorous proof of the $\Omega(\sqrt{H3SAT})$ regret bound for non-stationary MDPs. Finally, we discuss connections to PAC-MDP lower bounds.

Adversarially Guided Actor-Critic, [28]

Despite definite success in deep reinforcement learning problems, actor-critic algorithms are still confronted with sample inefficiency in complex environments, particularly in tasks where efficient exploration is a bottleneck. These methods consider a policy (the actor) and a value function (the critic) whose respective losses are built using different motivations and approaches. This paper introduces a third protagonist: the adversary. While the adversary mimics the actor by minimizing the KL-divergence between their respective action distributions, the actor, in addition to learning to solve the task, tries to differentiate itself from the adversary predictions. This novel objective stimulates the actor to follow strategies that could not have been correctly predicted from previous trajectories, making its behavior innovative in tasks where the reward is extremely rare. Our experimental analysis shows that the resulting Adversarially Guided Actor-Critic (AGAC) algorithm leads to more exhaustive exploration. Notably, AGAC outperforms current state-of-the-art methods on a set of various hard-exploration and procedurally-generated tasks.

SENTINEL: Taming Uncertainty with Ensemble-based Distributional Reinforcement Learning, [60]

In this paper, we consider risk-sensitive sequential decision-making in model-based reinforcement learning (RL). We introduce a novel quantification of risk, namely composite risk, which takes into account both aleatory and epistemic risk during the learning process.Previous works have considered aleatory or epistemic risk individually, or, an additive combination of the two. We demonstrate that the additive formulation is a particular case of the composite risk, which underestimates the actual CVaR risk even while learning a mixture of Gaussians. In contrast, the composite risk provides a more accurate estimate. We propose to use a bootstrapping method, SENTINEL-K, for distributional RL. SENTINEL-K uses an ensemble of K learners to estimate the return distribution and additionally uses follow the regularized leader (FTRL) from bandit literature for providing a better estimate of the risk on the return distribution. Finally, we experimentally verify that SENTINEL-K estimates the return distribution better, and while used with composite risk estimate,demonstrates better risk-sensitive performance than competing RL algorithms.

Demonstrating UDO: A Unified Approach for Optimizing Transaction Code, Physical Design, and System Parameters via Reinforcement Learning, [52]

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 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 configuration switching overheads. For instance, it aims to amortize the creation of expensive data structures by consecutively evaluating configurations using them. We demonstrate 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.

UDO: Universal Database Optimization using Reinforcement Learning, [53]

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.

8.3 Applications

Evaluating DAS3H on the EdNet Dataset, [20]

The EdNet dataset is a massive English language dataset that poses unique challenges for student performance prediction. In this paper, we describe and comment the results of our award-winning model DAS3H in the context of knowledge tracing in EdNet.

Deep Learning for Deep Waters: An Expert-in-the-Loop Machine Learning Framework for Marine Sciences, [15]

Driven by the unprecedented availability of data, machine learning has become a pervasive and transformative technology across industry and science. Its importance to marine science has been codified as one goal of the UN Ocean Decade. While increasing amounts of, for example, acoustic marine data are collected for research and monitoring purposes, and machine learning methods can achieve automatic

processing and analysis of acoustic data, they require large training datasets annotated or labelled by experts. Consequently, addressing the relative scarcity of labelled data is, besides increasing data analysis and processing capacities, one of the main thrust areas. One approach to address label scarcity is the expert-in-the-loop approach which allows analysis of limited and unbalanced data efficiently. Its advantages are demonstrated with our novel deep learning-based expert-in-the-loop framework for automatic detection of turbulent wake signatures in echo sounder data. Using machine learning algorithms, such as the one presented in this study, greatly increases the capacity to analyse large amounts of acoustic data. It would be a first step in realising the full potential of the increasing amount of acoustic data in marine sciences.

On Multi-Armed Bandit Designs for Dose-Finding Trials, [11]

We study the problem of finding the optimal dosage in early stage clinical trials through the multi-armed bandit lens. We advocate the use of the Thompson Sampling principle, a flexible algorithm that can accommodate different types of monotonicity assumptions on the toxicity and efficacy of the doses. For the simplest version of Thompson Sampling, based on a uniform prior distribution for each dose, we provide finite-time upper bounds on the number of sub-optimal dose selections, which is unprecedented for dose-finding algorithms. Through a large simulation study, we then show that variants of Thompson Sampling based on more sophisticated prior distributions outperform state-of-the-art dose identification algorithms in different types of dose-finding studies that occur in phase I or phase I/II trials.

8.4 Other

On Meritocracy in Optimal Set Selection, [58]

We consider the problem of selecting a set of individuals from a candidate population in order to maximise utility. When the utility function is defined over sets, this raises the question of how to define meritocracy. We define and analyse an appropriate notion of meritocracy derived from the utility function. We introduce the notion of expected marginal contributions of individuals and analyse its links to the underlying optimisation problem, our notion of meritocracy, and other notions of fairness such as the Shapley value. We also experimentally analyse the effect of different policy structures on the utility and meritocracy in a simulated college admission setting including constraints on statistical parity.

Stochastic Online Linear Regression: the Forward Algorithm to Replace Ridge, [40]

We consider the problem of online linear regression in the stochastic setting. We derive high probability regret bounds for online ridge regression and the forward algorithm. This enables us to compare online regression algorithms more accurately and eliminate assumptions of bounded observations and predictions. Our study advocates for the use of the forward algorithm in lieu of ridge due to its enhanced bounds and robustness to the regularization parameter. Moreover, we explain how to integrate it in algorithms involving linear function approximation to remove a boundedness assumption without deteriorating theoretical bounds. We showcase this modification in linear bandit settings where it yields improved regret bounds. Last, we provide numerical experiments to illustrate our results and endorse our intuitions.

Interferometric Graph Transform for Community Labeling, [61]

We present a new approach for learning unsupervised node representations in community graphs. We significantly extend the Interferometric Graph Transform (IGT) to community labeling: this non-linear operator iteratively extracts features that take advantage of the graph topology through demodulation operations. An unsupervised feature extraction step cascades modulus non-linearity with linear operators that aim at building relevant invariants for community labeling. Via a simplified model, we show that the IGT concentrates around the E-IGT: those two representations are related through some ergodicity properties. Experiments on community labeling tasks show that this unsupervised representation achieves performances at the level of the state of the art on the standard and challenging datasets Cora, Citeseer, Pubmed and WikiCS.

Low-Rank Projections of GCNs Laplacian, [33]

In this work, we study the behavior of standard models for community detection under spectral manipulations. Through various ablation experiments, we evaluate the impact of bandpass filtering on the performance of a GCN: we empirically show that most of the necessary and used information for nodes classification is contained in the low-frequency domain, and thus contrary to images, high

frequencies are less crucial to community detection. In particular, it is sometimes possible to obtain accuracies at a state-of-the-art level with simple classifiers that rely only on a few low frequencies.

Fast sampling from beta-ensembles, [13]

We study sampling algorithms for β-ensembles with time complexity less than cubic in the cardinality of the ensemble. Following Dumitriu & Edelman (2002), we see the ensemble as the eigenvalues of a random tridiagonal matrix, namely a random Jacobi matrix. First, we provide a unifying and elementary treatment of the tridiagonal models associated to the three classical Hermite, Laguerre and Jacobi ensembles. For this purpose, we use simple changes of variables between successive reparametrizations of the coefficients defining the tridiagonal matrix. Second, we derive an approximate sampler for the simulation of β-ensembles, and illustrate how fast it can be for polynomial potentials. This method combines a Gibbs sampler on Jacobi matrices and the diagonalization of these matrices. In practice, even for large ensembles, only a few Gibbs passes suffice for the marginal distribution of the eigenvalues to fit the expected theoretical distribution. When the conditionals in the Gibbs sampler can be simulated exactly, the same fast empirical convergence is observed for the fluctuations of the largest eigenvalue. Our experimental results support a conjecture by Krishnapur et al. (2016), that the Gibbs chain on Jacobi matrices of size N mixes in $\mathcal{O}(\log(N))$.

Justicia: A Stochastic SAT Approach to Formally Verify Fairness, [30]

As a technology ML is oblivious to societal good or bad, and thus, the field of fair machine learning has stepped up to propose multiple mathematical definitions, algorithms, and systems to ensure different notions of fairness in ML applications. Given the multitude of propositions, it has become imperative to formally verify the fairness metrics satisfied by different algorithms on different datasets. In this paper, we propose a stochastic satisfiability (SSAT) framework, Justicia, that formally verifies different fairness measures of supervised learning algorithms with respect to the underlying data distribution. We instantiate Justicia on multiple classification and bias mitigation algorithms, and datasets to verify different fairness metrics, such as disparate impact, statistical parity, and equalized odds. Justicia is scalable, accurate, and operates on non-Boolean and compound sensitive attributes unlike existing distribution-based verifiers, such as FairSquare and VeriFair. Being distribution-based by design, Justicia is more robust than the verifiers, such as AIF360, that operate on specific test samples. We also theoretically bound the finite-sample error of the verified fairness measure.

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: Philippe Preux, Odalric-Ambrym Maillard, Émilie Kaufmann, Debabrota Basu.

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: Ph. Preux

Partners:

- Bihar Agricultural University, India

Inria contact: Ph. Preux

Summary: in the context of Scool projects related to sustainable agriculture, the goal of this partnership is to collect data to train our learning algorithms.

10.1.2 STIC/MATH/CLIMAT AmSud project

- project STIC AmSud EMISTRAL managed by Inria Chile.

10.1.3 Participation in other International Programs

- I. Trummer, Assistant Professor, Cornell University, USA.
- K. Meel, Assistant Professor, National University of Singapore, Singapore.
- A. Schilep, Associate Professor, University of Gothenburg, Sweden.

10.2 International research visitors

Bishwamittra Ghosh

Status PhD

Institution of origin: National University of Singapore

Country: Singapore

Dates: Nov 2021- Feb 2022

Context of the visit: Bishwamittra Ghosh is a fourth-year PhD student at the Department of Computer Science in National University of Singapore. He is working at the intersection of machine learning and formal methods with Dr. Kuldeep S. Meel. His PhD research is on interpretable and fair machine learning (ML). He builds AI tools to learn interpretable ML models and to verify the fairness of ML models by relying on MaxSAT (maximum satisfiability), SSAT (stochastic satisfiability), etc. He is visiting Debabrota Basu at Scool. They have been collaborating on verification and explanation of unfairness in ML algorithms. During his visit, he plans to design a framework that can explain the source of unfairness by an ML model. Intuitively, the framework would compute an unfairness weight to an individual feature (or a subset of features) so that feature(s) with a higher weight is considered as the main source of unfairness induced by the model. Such weights are also defined as Fairness Influence Functions (FIF) of input features. There has been partial progress in computing FIF through borrowing techniques from explainability methods of ML models, such as based on Shapely values. In this research visit, Bishwamittra plans to further improve that direction.

Mobility program/type of mobility: Internship funded by MOBLILLEX scholarship of Université de Lille

10.2.1 Visits to international teams

Research stays abroad

Debabrota Basu

Visited institution: University of St. Gallen

Country: Switzerland

Dates: Nov 4-8, 2021

Context of the visit: D. Basu has visited the University of St. Gallen to work with his collaborator Prof. Christos Dimitrakakis. D. Basu has collaborating with C. Dimitrakakis and his students on multiple topics of reinforcement learning theory and fairness in sequential decision making. This visit aimed to conclude some of the ongoing projects and to begin new ones.

Mobility program/type of mobility: Research stay

10.3 European initiatives

10.3.1 Other european programs/initiatives

- Chist-Era project DELTA has been extended due to Covid and will end in 2022.
- Chist-Era project CausalXRL has started in 2021.

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 learnng for informationretrieval.
- É. Oyallon, CR CNRS, Sorbonne Université, machine learning on graphs.
- 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
- R. Rouvoy, Professor, Université de Lille, Inria Lille- Nord Europe (Équipe Spirals)
- A. Bellet, CR, Inria Lille- Nord Europe (Équipe Magnet)

10.5 Regional initiatives

- O.-A. Maillard and Ph. Preux co-chair an AI chair, funded by MEL, and Inria.
- Ph. Preux leads a collaboration with Prof. F. Pattou's service at CHU Lille/INSERM/Université de Lille regarding post-surgery patient follow-up. On Inria side, this collaboration involves an engineer (J. Teigny), and a Ph.D. Student (P. Saux). This collaboration is currently funded by a set of three regional projects.

11 Dissemination

11.1 Promoting scientific activities

11.1.1 Scientific events: selection

Member of the conference program committees

- Ph. Preux is in the PC of AAAI, ECML.
- E. Kaufmann is in the PC of ALT.
- O-A. Maillard is in the PC of NeurIPS (Area Chair)
- D. Basu is in the PC of AAAI, IJCAI, PoPETS.

Reviewer

- O-A. Maillard is reviewer at ICML, COLT and “emergency reviewer" at ICML.
- R. Degenne is reviewer at COLT, ALT and AISTATS.
- D. Basu is reviewer in the ICML, NeurIPS, AISTATS, and ICLR.

11.1.2 Journal

Member of the editorial boards

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

Reviewer - reviewing activities

- O-A. Maillard is reviewer for the journal Entropy, and for the Journal of the Royal Statistical Society.
- R. Degenne is reviewer for JMLR and the Journal on Uncertainty Quantification.
- D. Basu is reviewer for IEEE Access, IEEE Transactions on Information Forensics & Security (TIFS), IEEE Transactions on Dependable & Secure Computing (TDSC), Journal of American Statistical Association (JASA).

11.1.3 Invited talks

- Ph. Preux gives a talk entitled “Sequential decision making under uncertainty” at Digital Roads, Chile (virtual)
- E. Kaufmann gives a talk entitled “On pure exploration in (episodic) Markov Decision Processes” at the ICML workshop on Reinforcement Learning Theory (virtual).
- E. Kaufmann gives a talk entitled “Non-Parametric Exploration in Multi-Armed Bandits” at the Mathematical Statistics and Learning Worshop in Banff International Research Station (virtual).
- O-A. Maillard gives a talk entitled “Reinforcement Learning & Bandits for Agroecology” at the JFPDA (within PFIA, virtual).
- O-A. Maillard gives a talk entitled “A tour of Reinforcement Learning” at the Spiral team seminar, Inria Lille.
- O-A. Maillard gives a talk entitled “Some recent results in Reinforcement Learning theory” at the IMAG probability and statistics seminar, Université de Montpellier.
- D. Basu gives a talk entitled “SENTINEL: Quantifying Composite Uncertainty and Its Application in Risk-sensitive Reinforcement Learning" at the Learning Machines Seminars, Research Institute of Sweden (RISE) (virtual).
- D. Basu gives a series of talks entitled “Privacy Preserving Machine Learning" at the Indian Statistical Institute (ISI) Kolkata (virtual).

11.1.4 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 Inria DR 2 competition,
  - a member of the CRCN competition in Inria-Rennes,
  - a member of the airlab jury at Université de Lille.

11.1.5 Research administration

- Ph. Preux is deputy scientific delegate at Inria Lille.

11.2 Teaching - Supervision - Juries

11.2.1 Teaching

- D. Basu: “Research Reading Group", M2 Data Science, Centrale Lille in 2021.
- D. Basu: “Anonymity and Privacy", M2 Artificial Intelligence and Data Science, École Normale Superieure (ENS)- PSL University in 2021.
- D. Baudry taught about 64 h. at Université de Lille in 2021, in maths (L2 MIAH).
- O. Darwiche taught reinforcement learning at École Centrale de Lille (3 practical sessions) and he is grading reinforcement learning homeworks for the MVA Master at ENS Cachan.
- E. Kaufmann: Reinforcement Learning (20h), Ecole Centrale Lille.
- O-A. Maillard: Statistical Reinforcement Learning (42h), MAP/INF641, Master Artificial Intelligence and advanced Visual Computing, École Polytechnique.

- O-A. Maillard: Reinforcement Learning (20h), Master 2 Artificial Intelligence, École CentraleSupélec.
- F. Pesquerel: TA for Ph. Preux “prise de décision séquentielle dans l’incertain”, M2 in Comptuter Science, Université de Lille
- F. Pesquerel: TA for O-A. Maillard “Reinforcement Learning”, M2 in Computer Science, École CentraleSupélec
- F. Pesquerel: TA for O-A. Maillard “statistical Reinforcement Learning”, MAP/INF641, Master Artificial Intelligence and advanced Visual Computing, École Polytechnique.
- P. Saux: TA for O-A. Maillard “Reinforcement Learning”, M2 in Comptuter Science, École Centrale-Supélec
- Ph. Preux: « IA et apprentissage automatique », DU IA & Santé, Université de Lille
- Ph. Preux: prise de décision séquentielle dans l’incertain, M2 in Comptuter Science, Université de Lille
- R. Degenne: “Sequential learning”, M2 MVA, ENS Paris-Saclay

11.2.2 Supervision

- Ph. Preux supervised the internship of:
  - Matheus Medeiros Centa, M2 Maths, Sorbonne Université, Paris,
  - Danilo Moreira Lagos, L3 Computer Science, Chile,
  - Toby Johnstone, MVA, ENS Paris-Saclay, co-supervised with N. Grinzstajn.
- O-A. Maillard supervised the internship of:
  - Thomas Carta, M2 Maths, École Polytechnique, Saclay.
- R. Degenne supervised the internship of:
  - Marc Jourdan, Master, ETH Zurich
- D. Basu supervised the internship of:
  - Pratik Karmakar, Master, RKMVERI, India

11.2.3 Juries

- Ph. Preux was a member of the juries of:
  - Ph.D. in medicine of Clémence Rozwag, Université de Lille
  - Ph.D. in CS of Mathieu Seurin, Université de Lille
  - Ph.D. in CS of Yannis Flet-Berliac, Université de Lille
  - Ph.D. in CS of Jérôme Buisine, Université du Littoral Côte d’Opale
- E. Kaufmann was a member of the juries of :
  - Ph.D. in CS of Louis Faury (Telecom ParisTech) (reviewer)
  - Ph.D. in CS of Xuedong Shang, Université de Lille (PhD advisor)
  - Ph.D. in maths of Zixin Zong, National University of Singapore (reviewer)
  - Ph.D. in CS of Mathiey Chambefort, Mines ParisTech
- O-A. Maillard was a member of the juries of :
  - Ph.D. in CS of Réda Alami, CIFRE, Université Paris-Saclay (PhD co-advisor)
  - Ph.D. in CS of Thibaut Cuvelier, CIFRE, CentraleSupélec (reviewer)
  - Ph.D. CST in CS of Houssam Zenati, CIFRE, Univ. Grenoble (reviewer)
  - INRAE concours de chargé-e-s de recherche de classe normale (MathNum department)

11.3 Popularization

11.3.1 Interventions

- Ph. Preux gives a talk on “Artificial Intelligence” at the “ethics and AI” organized at the Université de Lille.

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] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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

[11] M. Aziz, E. Kaufmann and M.-K. Riviere. 'On Multi-Armed Bandit Designs for Dose-Finding Trials'. In: *Journal of Machine Learning Research* (Jan. 2021). URL: https://hal.archives-ouvertes.fr/hal-02533297.

[12] A. Garivier and E. Kaufmann. 'Non-Asymptotic Sequential Tests for Overlapping Hypotheses and application to near optimal arm identification in bandit models'. In: *Sequential Analysis* (Mar. 2021). URL: https://hal.archives-ouvertes.fr/hal-02123833.

[13] G. Gautier, R. Bardenet and M. Valko. 'Fast sampling from beta-ensembles'. In: *Statistics and Computing* 31.7 (12th Jan. 2021). DOI: 10.1007/s11222-020-09984-0. URL: https://hal.archives-ouvertes.fr/hal-02697647.

[14] E. Kaufmann and W. M. Koolen. 'Mixture Martingales Revisited with Applications to Sequential Tests and Confidence Intervals'. In: *Journal of Machine Learning Research* (6th Dec. 2021). URL: https://hal.archives-ouvertes.fr/hal-01886612.

[15] I. Ryazanov, A. Nylund, D. Basu, I.-M. Hassellöv and A. Schliep. 'Deep Learning for Deep Waters: An Expert-in-the-Loop Machine Learning Framework for Marine Sciences'. In: *Journal of Marine Science and Engineering* 9.2 (Feb. 2021), p. 169. DOI: 10.3390/jmse9020169. URL: https://hal.archives-ouvertes.fr/hal-03445756.

International peer-reviewed conferences

[16] M. Andrychowicz, A. Raichuk, P. Stańczyk, M. Orsini, S. Girgin, R. Marinier, L. Hussenot, M. Geist, O. Pietquin, M. Michalski, S. Gelly and O. Bachem. 'What Matters In On-Policy Reinforcement Learning? A Large-Scale Empirical Study'. In: ICLR 2021 - Ninth International Conference on Learning Representations. Vienna / Virtual, Austria, 4th May 2021. URL: https://hal.inria.fr/hal-03162554.

[17] 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.archives-ouvertes.fr/hal-03447244.

[18] D. Baudry, Y. Russac and O. Cappé. 'On Limited-Memory Subsampling Strategies for Bandits'. In: ICML 2021- International Conference on Machine Learning. Vienna / Virtual, Austria, 18th July 2021. URL: https://hal.archives-ouvertes.fr/hal-03265442.

[19] D. Baudry, P. Saux and O.-A. Maillard. 'From Optimality to Robustness: Dirichlet Sampling Strategies in Stochastic Bandits'. In: NeurIPS 2021 - 35th International Conference on Neural Information Processing Systems. Sydney, Australia, 6th Dec. 2021. URL: https://hal.archives-ouvertes.fr/hal-03421252.

[20] B. Choffin, F. Popineau, Y. Bourda and J.-J. Vie. 'Evaluating DAS3H on the EdNet Dataset'. In: AAAI 2021 - The 35th Conference on Artificial Intelligence / Imagining Post-COVID Education with AI. Virtual, United States, 20th Jan. 2021. URL: https://hal.archives-ouvertes.fr/hal-03175874.

[21] S. R. Chowdhury, A. Gopalan and O.-A. Maillard. 'Reinforcement Learning in Parametric MDPs with Exponential Families'. In: *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*. International Conference on Artificial Intelligence and Statistics. Vol. 130. Proceedings of Machine Learning Research. San diego, United States, 2021, pp. 1855–1863. URL: https://hal.archives-ouvertes.fr/hal-03472116.

[22] R. Dadashi, L. Hussenot, M. Geist and O. Pietquin. 'Primal Wasserstein Imitation Learning'. In: ICLR 2021 - Ninth International Conference on Learning Representations. Vienna / Virtual, Austria, 8th June 2020. URL: https://hal.inria.fr/hal-03162526.

[23] R. Dadashi, S. Rezaeifar, N. Vieillard, L. Hussenot, O. Pietquin and M. Geist. 'Offline Reinforcement Learning with Pseudometric Learning'. In: ICML 2021 - 38th International Conference on Machine Learning. Vol. 139. virtual, France, 18th June 2021. URL: https://hal.inria.fr/hal-03468847.

[24] O. D. Domingues, P. Ménard, M. Pirotta, E. Kaufmann and M. Valko. 'A kernel-based approach to non-stationary reinforcement learning in metric spaces'. In: International Conference on Artificial Intelligence and Statistics. San Diego / Virtual, United States, 13th Apr. 2021. URL: https://hal.inria.fr/hal-03289026.

[25] O. D. Domingues, P. Ménard, M. Pirotta, E. Kaufmann and M. Valko. 'Kernel-based reinforcement Learning: A finite-time analysis'. In: International Conference on Machine Learning. Vienna / Virtual, Austria, 18th July 2021. URL: https://hal.inria.fr/hal-02541790.

[26] O. D. Domingues, P. Ménard, E. Kaufmann and M. Valko. 'Episodic reinforcement learning in finite MDPs: Minimax lower bounds revisited'. In: Algorithmic Learning Theory. Paris / Virtual, France, 16th Mar. 2021. URL: https://hal.inria.fr/hal-03289004.

[27] J. Ferret, O. Pietquin and M. Geist. 'Self-Imitation Advantage Learning'. In: AAMAS 2021 - 20th International Conference on Autonomous Agents and Multiagent Systems. Londres / Virtual, United Kingdom, 3rd May 2021. URL: https://hal.inria.fr/hal-03159815.

[28] Y. Flet-Berliac, J. Ferret, O. Pietquin, P. Preux and M. Geist. 'Adversarially Guided Actor-Critic'. In: ICLR 2021 - International Conference on Learning Representations. Vienna / Virtual, Austria, 4th May 2021. URL: https://hal.inria.fr/hal-03167169.

[29] Y. Flet-Berliac, R. Ouhamma, O.-A. Maillard and P. Preux. 'Learning Value Functions in Deep Policy Gradients using Residual Variance'. In: ICLR 2021 - International Conference on Learning Representations. Vienna / Virtual, Austria, 4th May 2021. URL: https://hal.archives-ouvertes.fr/hal-02964174.

[30] 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.archives-ouvertes.fr/hal-03445831.

[31] N. Grinsztajn, O. Beaumont, E. Jeannot and P. Preux. 'READYS: A Reinforcement Learning Based Strategy for Heterogeneous Dynamic Scheduling'. In: IEEE Cluster 2021. Portland / Virtual, United States, 7th Sept. 2021. URL: https://hal.inria.fr/hal-03313229.

[32] N. Grinsztajn, J. Ferret, O. Pietquin, P. Preux and M. Geist. 'There Is No Turning Back: A Self-Supervised Approach for Reversibility-Aware Reinforcement Learning'. In: *Proc. Thirty-fifth Conference on Neural Information Processing Systems*. Neural Information Processing Systems (2021). Virtual, France, 6th Dec. 2021. URL: https://hal.archives-ouvertes.fr/hal-03454640.

[33] N. Grinsztajn, P. Preux and E. Oyallon. 'Low-Rank Projections of GCNs Laplacian'. In: ICLR 2021 Workshop GTRL. Online, France, 7th May 2021. URL: https://hal.archives-ouvertes.fr/hal-03248056.

[34] L. Hussenot, R. Dadashi, M. Geist and O. Pietquin. 'Show me the Way: Intrinsic Motivation from Demonstrations'. In: AAMAS 2021 - 20th International Conference on Autonomous Agents and Multiagent Systems. Virtual, United Kingdom, 3rd May 2021. URL: https://hal.inria.fr/hal-03162139.

[35] 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.archives-ouvertes.fr/hal-02864574.

[36] P. Ménard, O. D. Domingues, E. Kaufmann, A. Jonsson, E. Leurent and M. Valko. 'Fast active learning for pure exploration in reinforcement learning'. In: International Conference on Machine Learning. Vienna, Austria, 18th July 2021. URL: https://hal.inria.fr/hal-02906985.

[37] P. Ménard, O. D. Domingues, X. Shang and M. Valko. 'UCB Momentum Q-learning: Correcting the bias without forgetting'. In: International Conference on Machine Learning. Vienna / Virtual, Austria, 18th July 2021. URL: https://hal.inria.fr/hal-03289033.

[38] M. H. Nguyen, N. Grinsztajn, I. Guyon and L. Sun-Hosoya. 'MetaREVEAL: RL-based Meta-learning from Learning Curves'. In: Workshop on Interactive Adaptive Learning co-located with European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD 2021). Bilbao/Virtual, Spain, 13th Sept. 2021. URL: https://hal.inria.fr/hal-03502358.

[39] R. Ouhamma, R. Degenne, P. Gaillard and V. Perchet. 'Online Sign Identification: Minimization of the Number of Errors in Thresholding Bandits'. In: NeurIPS 2021 - 35th International Conference on Neural Information Processing Systems. NeurIPS 2021 - 35th International Conference on Neural Information Processing Systems. Virtual, Canada, 2021, pp. 1–25. URL: https://hal.inria.fr/hal-03363014.

[40] R. Ouhamma, O. Maillard and V. Perchet. 'Stochastic Online Linear Regression: the Forward Algorithm to Replace Ridge'. In: NeurIPS 2021 - 35th International Conference on Neural Information Processing Systems. NeurIPS 2021 - 35th International Conference on Neural Information Processing Systems. Virtual, Canada, 6th Dec. 2021. URL: https://hal.inria.fr/hal-03410901.

[41] M. Papini, A. Tirinzoni, A. Pacchiano, M. Restilli, A. Lazaric and M. Pirotta. 'Reinforcement Learning in Linear MDPs: Constant Regret and Representation Selection'. In: Thirty-Fifth Conference on Neural Information Processing Systems. Virtual, France, 6th Dec. 2021. URL: https://hal.inria.fr/hal-03479324.

[42] F. Pesquerel, H. Saber and O.-A. Maillard. 'Stochastic bandits with groups of similar arms'. In: *NeurIPS 2021 - Thirty-fifth Conference on Neural Information Processing Systems*. NeurIPS 2021 - Thirty-fifth Conference on Neural Information Processing Systems. Stochastic bandits with groups of similar arms. Sydney, Australia, 6th Dec. 2021. URL: https://hal.archives-ouvertes.fr/hal-03427597.

[43] C. Réda, E. Kaufmann and A. Delahaye-Duriez. 'Top-m identification for linear bandits'. In: Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS). Vol. 130. Virtual, United States, 2021. URL: https://hal.archives-ouvertes.fr/hal-03172145.

[44] C. Réda, A. Tirinzoni and R. Degenne. 'Dealing With Misspecification In Fixed-Confidence Linear Top-m Identification'. In: 35th Conference on Neural Information Processing Systems. Virtual, France, 2021. URL: https://hal.archives-ouvertes.fr/hal-03409205.

[45] 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.

[46] H. Saber, P. Ménard and O.-A. Maillard. 'Indexed Minimum Empirical Divergence for Unimodal Bandits'. In: *NeurIPS 2021 - International Conference on Neural Information Processing Systems*. NeurIPS 2021 - International Conference on Neural Information Processing Systems. Virtual-only Conference, United States, 6th Dec. 2021. URL: https://hal.archives-ouvertes.fr/hal-03446617.

[47] H. Saber, L. Saci, O.-A. Maillard and A. Durand. 'Routine Bandits: Minimizing Regret on Recurring Problems'. In: *ECML-PKDD 2021*. ECML-PKDD 2021. Bilbao, Spain, 13th Sept. 2021. URL: https://hal.archives-ouvertes.fr/hal-03286539.

[48] M. Seurin, F. Strub, P. Preux and O. Pietquin. 'Don't Do What Doesn't Matter: Intrinsic Motivation with Action Usefulness'. In: *Proc. Internationnal Joint Conference on Artificial Intelligence (IJCAI)*. Internationnal Joint Conference on Artificial Intelligence (IJCAI). Montreal, Canada, 21st Aug. 2021, pp. 2950–2956. URL: https://hal.archives-ouvertes.fr/hal-03259315.

[49] J. Tarbouriech, M. Pirotta, M. Valko and A. Lazaric. 'A Provably Efficient Sample Collection Strategy for Reinforcement Learning'. In: Neural Information Processing Systems (NeurIPS). Virtual/Sydney, Australia, 6th Dec. 2021. URL: https://hal.inria.fr/hal-03479827.

[50] J. Tarbouriech, M. Pirotta, M. Valko and A. Lazaric. 'Sample complexity bounds for stochastic shortest path with a generative model'. In: Algorithmic Learning Theory. Paris, France, 2021. URL: https://hal.inria.fr/hal-03288988.

[51] J. Tarbouriech, R. Zhou, S. S. Du, M. Pirotta, M. Valko and A. Lazaric. 'Stochastic Shortest Path: Minimax, Parameter-Free and Towards Horizon-Free Regret'. In: Neural Information Processing Systems (NeurIPS). Virtual/Sydney, Australia, 6th Dec. 2021. URL: https://hal.inria.fr/hal-03479782.

[52] J. Wang, I. Trummer and D. Basu. 'Demonstrating UDO: A Unified Approach for Optimizing Transaction Code, Physical Design, and System Parameters via Reinforcement Learning'. In: SIGMOD/PODS '21: International Conference on Management of Data. Proceedings of the 2021 International Conference on Management of Data (SIGMOD '21). Virtual Event, China: ACM, June 2021, pp. 2794–2797. DOI: 10.1145/3448016.3452754. URL: https://hal.archives-ouvertes.fr/hal-03446016.

[53] 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.

Doctoral dissertations and habilitation theses

[54] Y. Flet-Berliac. 'Sample-Efficient Deep Reinforcement Learning for Control, Exploration and Safety'. Université de Lille - Faculté des Sciences et Technologies, 6th Oct. 2021. URL: https://tel.archives-ouvertes.fr/tel-03431652.

[55] M. Seurin. 'Learning to Interact, Interacting to Learn Action-centric Reinforcement Learning'. Université de Lille - Faculté des Sciences et Technologies, 28th Sept. 2021. URL: https://tel.archives-ouvertes.fr/tel-03432794.

[56] X. Shang. 'Adaptive methods for optimization in stochastic environments'. Université de Lille, 29th Sept. 2021. URL: https://tel.archives-ouvertes.fr/tel-03466525.

Reports & preprints

[57] 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.

[58] T. K. Buening, M. Segal, D. Basu, C. Dimitrakakis and A.-M. George. *On Meritocracy in Optimal Set Selection*. 24th Nov. 2021. URL: https://hal.archives-ouvertes.fr/hal-03445971.

[59] T. Carta, S. Chaudhury, K. Talamadupula and M. Tatsubori. *VISUALHINTS: A Visual-Lingual Environment for Multimodal Reinforcement Learning*. 6th Dec. 2021. URL: https://hal.archives-ouvertes.fr/hal-03466647.

[60] H. Eriksson, D. Basu, M. Alibeigi and C. Dimitrakakis. *SENTINEL: Taming Uncertainty with Ensemble-based Distributional Reinforcement Learning*. 24th Feb. 2021. URL: https://hal.archives-ouvertes.fr/hal-03150823.

[61] N. Grinsztajn, L. Leconte, P. Preux and E. Oyallon. *Interferometric Graph Transform for Community Labeling*. 4th June 2021. URL: https://hal.archives-ouvertes.fr/hal-03247781.

[62] J. Wang, D. Basu and I. Trummer. *Procrastinated Tree Search: Black-box Optimization with Delayed, Noisy, and Multi-fidelity Feedback*. 24th Nov. 2021. URL: https://hal.archives-ouvertes.fr/hal-03445909.