RESEARCH CENTRE

Lille - Nord Europe

IN PARTNERSHIP WITH:

CNRS, Université de Lille

2020 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 July 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

B4.4. – Energy delivery

B4.4.1. – Smart grids

B5.8. – Learning and training

B7.2.1. – Smart vehicles

B9.1.1. – E-learning, MOOC

B9.5. – Sciences

B9.5.6. – Data science

1 Team members, visitors, external collaborators

Research Scientists

- Debabrota Basu [Inria, from Nov 2020, Starting Faculty Position]
- Rémy Degenne [Inria, from Nov 2020, Starting Faculty Position]
- Émilie Kaufmann [CNRS, Researcher, HDR]
- Odalric-Ambrym Maillard [Inria, Researcher, HDR]
- Jill-Jênn Vie [Inria, Researcher]

Faculty Member

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

Post-Doctoral Fellows

- Sein Minn [Inria, from Aug 2020]
- Mohit Mittal [Inria, from Nov 2020]
- Pierre Ménard [Inria, until Oct 2020]

PhD Students

- Dorian Baudry [CNRS]
- Omar Darwiche Domingues [Inria]
- Johan Ferret [Google]
- Yannis Flet-Berliac [Université de Lille]
- Guillaume Gautier [CNRS, Until Oct 2020]
- Nathan Grinsztajn [École polytechnique]
- Leonard Hussenot-Desenonges [Google]
- Édouard Leurent [Renault, Until Oct 2020]
- Reda Ouhamma [École polytechnique]
- Pierre Perrault [Inria, until Nov 2020]
- Sarah Perrin [Université de Lille]
- Fabien Pesquerel [École Normale Supérieure de Paris, from Nov 2020]
- Clémence Réda [Université de Paris]
- Hassan Saber [Inria, until Aug 2020]
- Patrick Saux [Inria, from Nov 2020]
- Mathieu Seurin [Université de Lille]
- Julien Seznec [Le Livre Scolaire, until Dec 2020]
- Xuedong Shang [Université de Lille]
- Florian Strub [Deepmind, until Jan 2020]
- Jean Tarbouriech [Facebook]

Technical Staff

- Clémence Léguillette [Inria, Engineer, from Oct 2020]
- Vianney Taquet [Inria, Engineer, from Sep 2020]
- Julien Teigny [Inria, Engineer, from Sep 2020]
- Franck Valentini [Inria, Engineer, until Oct 2020]

Administrative Assistant

- Amélie Supervielle [Inria]

External Collaborator

- Romain Gautron [Centre de coopération internationale en recherche agronomique]

2 Overall objectives

Scool is a machine learning (ML) research group. Scool 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 let 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 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
- RL in non stationary environments

Beyond these objectives, we put a particular emphasis on the study of non-stationary environments. An other 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 3 main topics of application:

- health
- sustainable development
- e-learning

In each of these 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 been yet extensively studied and applied in these fields, sequential decision making remains far less studied; bandits have yet been used in many applications of e-commerce (e.g. for computational advertising and recommendation systems). However, in applications where human beings may be severaly impacted, bandits and reinforcement learning have not yet been much studied; moreover, these applications come along 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 2020, 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 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). In 2019-2020, we also studied more traditional machine learning aspects, with the investigation of deep learning technology in radiology, with Prof. A. Cotten (CHU and Université de Lille), in project RAID. Finally, in the context of the Covid-19 sudden pandemic, we volunteered to get involved in a series of works with AP-HP; if the immediate needs were not on research question but more about the exploitation of our skills in data science and machine learning, this activity led towards more research oriented questions later in Fall 2020.

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. In marine biology, we focus on the warm water footprints, called wakes, caused by movements of large marine vessels. These wakes disrupt the natural flora and fauna around the ship routes and ports. We are developing machine learning techniques to detect the wakes using data from acoustic Doppler profilers, and to correlate them with the structure and motion of marine vessels. We also worked on the control of smartgrids, with LPCIM at École Polytechnique, and in collaboration with Total.

Regarding e-learning, the collaboration with Le Livre Scolaire has led to the defense of J. Seznec's PhD (CIFRE grant). We have continued our work with Pix by organizing an open lab in January 2020 attracting 20 people of various fields: psychometricians, data scientists, statisticians at the French Ministry of Education. We also got involved in advising a start-up under creation related to e-learning, this creation being supported by the Inria Startup Studio.

There are two important aspects that are amply shared common by these various 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. This comes along the work we did on the autonomous control of vehicles (with Renault, in collaboration with Inria-Lille team Valse, through a CIFRE grant) where safety is obviously a very important issue too.

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 locks, and styudying how to overcome them.

Let us mention here:

- wake detection in marine sciences,
- sustainable agriculture in developing countries,
- sustainable gardening,
- control of smartgrids.

More details can be found in section 4.

6 Highlights of the year

- On Oct. 31st 2020, SequeL ended, being replaced by the brand new joint team-project Scool.
- É. Leurent had an oral presentation at NeurIPS (1% acceptance rate this year). This magnifically concludes his PhD work based on a collaboration between SequeL/Scool and Valse, under O.-A. Maillard and D. Efimov co-supervision.
- Our team composed of J.-J. Vie, Sein Minn (Scool), Mehdi Douch, Yassine Esmili (Axiome, Inria Startup Studio) got ranked 5 over 23 submissions at the NeurIPS Education Challenge 2020. J.-J. Vie (Scool) and Matthieu Doutreligne (Parietal, Inria Saclay) got ranked 2 over 16 submissions at the NeurIPS Healthcare Hide-and-Seek Challenge 2020.
- The Inria-Covid ScikitEDS project in partnership with Inria Parietal was presented to President Emmanuel Macron on Dec. 4, 2020.

7 New software and platforms

7.1 New software

7.1.1 highway-env

Name: An environment for autonomous driving decision-making

Keywords: Generic modeling environment, Simulation, Autonomous Cars, Artificial intelligence

Functional Description: The environment is composed of several variants, each of which corresponds to driving scenes: highway, roundabout, intersection, merge, parking, etc. The road network is described by a graph, and is then populated with simulated vehicles. Vehicle kinematics follows a simple Bicycle model, and their behavior is determined by models derived from road traffic simulation literature. The ego-vehicle has access to a description of the scene through several types of observations, and its behavior is controlled through an action space, either discrete (change of lanes, of cruising speed) or continuous (accelerator pedal, steering wheel angle). The objective function to maximize is also described by the environment and may vary depending on the task to be solved. The interface of the library is inherited from the standard defined by OpenAI Gym, consisting of four main methods: gym.make(id), env.step(action), env.reset(), and env.render().

URL: https://github.com/eleurent/highway-env

Author: Edouard Leurent

Contact: Edouard Leurent

7.1.2 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

7.1.3 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

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

Statistical efficiency of Thompson sampling for combinatorial semi-bandits [40]

We investigate stochastic combinatorial multi-armed bandit with semi-bandit feedback (CMAB). In CMAB, the question of the existence of an efficient policy with an optimal asymptotic regret (up to a factor poly-logarithmic with the action size) is still open for many families of distributions, including mutually independent outcomes, and more generally the multivariate sub-Gaussian family. We propose to answer the above question for these two families by analyzing variants of the Combinatorial Thompson Sampling policy (CTS). For mutually independent outcomes in $[0,1]$, we propose a tight analysis of CTS using Beta priors. We then look at the more general setting of multivariate sub-Gaussian outcomes and propose a tight analysis of CTS using Gaussian priors. This last result gives us an alternative to the Efficient Sampling for Combinatorial Bandit policy (ESCB), which, although optimal, is not computationally efficient.

Sub-sampling for Efficient Non-Parametric Bandit Exploration [19]

In this paper we propose the first multi-armed bandit algorithm based on re-sampling that achieves asymptotically optimal regret simultaneously for different families of arms (namely Bernoulli, Gaussian and Poisson distributions). Unlike Thompson Sampling which requires to specify a different prior to be optimal in each case, our proposal RB-SDA does not need any distribution-dependent tuning. RB-SDA belongs to the family of Sub-sampling Duelling Algorithms (SDA) which combines the sub-sampling idea first used by the BESA [61] and SSMC [63] algorithms with different sub-sampling schemes. In particular, RB-SDA uses Random Block sampling. We perform an experimental study assessing the flexibility and robustness of this promising novel approach for exploration in bandit models.

Monte-Carlo Graph Search: the Value of Merging Similar States [38]

We consider the problem of planning in a Markov Decision Process (MDP) with a generative model and limited computational budget. Despite the underlying MDP transitions having a graph structure, the popular Monte-Carlo Tree Search algorithms such as UCT rely on a tree structure to represent their value estimates. That is, they do not identify together two similar states reached via different trajectories and represented in separate branches of the tree. In this work, we propose a graph-based planning algorithm, which takes into account this state similarity. In our analysis, we provide a regret bound that depends on a novel problem-dependent measure of difficulty, which improves on the original tree-based bound in MDPs where the trajectories overlap, and recovers it otherwise. Then, we show that this methodology can be adapted to existing planning algorithms that deal with stochastic systems. Finally, numerical simulations illustrate the benefits of our approach.

Planning in Markov Decision Processes with Gap-Dependent Sample Complexity [33]

We propose MDP-GapE, a new trajectory-based Monte-Carlo Tree Search algorithm for planning in a Markov Decision Process in which transitions have a finite support. We prove an upper bound on the number of calls to the generative models needed for MDP-GapE to identify a near-optimal action with high probability. This problem-dependent sample complexity result is expressed in terms of the sub-optimality gaps of the state-action pairs that are visited during exploration. Our experiments reveal that MDP-GapE is also effective in practice, in contrast with other algorithms with sample complexity guarantees in the fixed-confidence setting, that are mostly theoretical.

Spectral bandits [13]

Smooth functions on graphs have wide applications in manifold and semi-supervised learning. In this work, we study a bandit problem where the payoffs of arms are smooth on a graph. This framework is suitable for solving online learning problems that involve graphs, such as content-based recommendation. In this problem, each item we can recommend is a node of an undirected graph and its expected rating is

similar to the one of its neighbors. The goal is to recommend items that have high expected ratings. We aim for the algorithms where the cumulative regret with respect to the optimal policy would not scale poorly with the number of nodes. In particular, we introduce the notion of an effective dimension, which is small in real-world graphs, and propose three algorithms for solving our problem that scale linearly and sublinearly in this dimension. Our experiments on contentrecommendation problem show that a good estimator of user preferences for thousands of items can be learned from just tens of node evaluations.

Solving Bernoulli Rank-One Bandits with Unimodal Thompson Sampling [46]

Stochastic Rank-One Bandits [68, 69] are a simple framework for regret minimization problems over rank-one matrices of arms. The initially proposed algorithms are proved to have logarithmic regret, but do not match the existing lower bound for this problem. We close this gap by first proving that rank-one bandits are a particular instance of unimodal bandits, and then providing a new analysis of Unimodal Thompson Sampling (UTS), initially proposed by Paladino et al. [70]. We prove an asymptotically optimal regret bound on the frequentist regret of UTS and we support our claims with simulations showing the significant improvement of our method compared to the state-of-the-art.

A Practical Algorithm for Multiplayer Bandits when Arm Means Vary Among Players [20]

We study a multiplayer stochastic multi-armed bandit problem in which players cannot communicate, and if two or more players pull the same arm,a collision occurs and the involved players receive zero reward. We consider the challenging heterogeneous setting, in which different arms may have different means for different players, and propose a new and efficient algorithm that combines the idea of leveraging forced collisions for implicitcommunication and that of performing matching eliminations. We present a finite-time analysis of our algorithm, giving the first sublinear minimaxregret bound for this problem, and prove that if the optimal assignment of players to arms is unique, our algorithm attains the optimal $O(\ln(T))$ regret, solving an open question raised at NeurIPS 2018 by Bistritz and Leshem [62].

Budgeted online influence maximization [41]

We introduce a new budgeted framework for on-line influence maximization, considering the total cost of an advertising campaign instead of the common cardinality constraint on a chosen influ-encer set. Our approach models better the real-world setting where the cost of influencers varies and advertizers want to find the best value for their overall social advertising budget. We propose an algorithm assuming an independent cascade diffusion model and edge-level semi-bandit feedback, and provide both theoretical and experimental results. Our analysis is also valid for the cardinality-constraint setting and improves the state of the art regret bound in this case.

Fixed-confidence guarantees for Bayesian best-arm identification [45]

We investigate and provide new insights on the sampling rule called Top-Two Thompson Sampling (TTTS). In particular, we justify its use for fixed-confidence best-arm identification. We further propose a variant of TTTS called Top-Two Transportation Cost (T3C), which disposes of the computational burden of TTTS. As our main contribution, we provide the first sample complexity analysis of TTTS and T3C when coupled with a very natural Bayesian stopping rule, for bandits with Gaussian rewards, solving one of the open questions raised by Russo [71]. We also provide new posterior convergence results for TTTS under two models that are commonly used in practice: bandits with Gaussian and Bernoulli rewards and conjugate priors.

The Influence of Shape Constraints on the Thresholding Bandit Problem [21]

We investigate the stochastic Thresholding Bandit problem (TBP) under several shape constraints. On top of (i) the vanilla, unstructured TBP, we consider the case where (ii) the sequence of arm's means $(\mu_k)_k$ is monotonically increasing MTBP, (iii) the case where $(\mu_k)_k$ is unimodal UTBP and (iv) the case where $(\mu_k)_k$ is concave CTBP. In the TBP problem the aim is to output, at the end of the sequential game, the set of arms whose means are above a given threshold. The regret is the highest gap between a misclassified arm and the threshold. In the fixed budget setting, we provide problem independent minimax rates for the expected regret in all settings, as well as associated algorithms. We prove that the minimax rates for the regret are (i) $\sqrt{\log(K)K/T}$ for TBP, (ii) $\sqrt{\log(K)/T}$ for MTBP, (iii) $\sqrt{K/T}$ for

UTBP and (iv) $\sqrt{\log\log K/T}$ for CTBP, where K is the number of arms and T is the budget. These rates demonstrate that the dependence on K of the minimax regret varies significantly depending on the shape constraint. This highlights the fact that the shape constraints modify fundamentally the nature of the TBP problem to the other.

Covariance-adapting algorithm for semi-bandits with application to sparse outcomes [42]

We investigate stochastic combinatorial semi-bandits, where the entire joint distribution of outcomes impacts the complexity of the problem instance (unlike in the standard bandits). Typical distributions considered depend on specific parameter values, whose prior knowledge is required in theory but quite difficult to estimate in practice; an example is the commonly assumed sub-Gaussian family. We alleviate this issue by instead considering a new general family of sub-exponential distributions, which contains bounded and Gaussian ones. We prove a new lower bound on the regret on this family, that is parameterized by the unknown covariance matrix, a tighter quantity than the sub-Gaussian matrix. We then construct an algorithm that uses covariance estimates, and provide a tight asymptotic analysis of the regret. Finally, we apply and extend our results to the family of sparse outcomes, which has applications in many recommender systems.

Gamification of pure exploration for linear bandits [24]

We investigate an active pure-exploration setting, that includes best-arm identification, in the context of linear stochastic bandits. While asymptotically optimal algorithms exist for standard multi-arm bandits, the existence of such algorithms for the best-arm identification in linear bandits has been elusive despite several attempts to address it. First, we provide a thorough comparison and new insight over different notions of optimality in the linear case, including G-optimality, transductive optimality from optimal experimental designand asymptotic optimality. Second, we design the first asymptotically optimal algorithm for fixed-confidence pure exploration in linear bandits. As a consequence, our algorithm naturally bypasses the pitfall caused by a simple but difficult instance, that most prior algorithms had to be engineered to deal with explicitly. Finally, we avoid the need to fully solve an optimal design problem by providing an approach that entails an efficient implementation.

Stochastic bandits with vector losses: Minimizing ℓ^∞-norm of relative losses [60]

Multi-armed bandits are widely applied in scenarios like recommender systems, for which the goal is to maximize the click rate. However, more factors should be considered, e.g., user stickiness, user growth rate, user experience assessment, etc. In this paper, we model this situation as a problem of K-armed bandit with multiple losses. We define relative loss vector of an arm where the i-th entry compares the arm and the optimal arm with respect to the i-th loss. We study two goals: (a) finding the arm with the minimum ℓ^∞-norm of relative losses with a given confidence level (which refers to fixed-confidence best-arm identification); (b) minimizing the ℓ^∞-norm of cumulative relative losses (which refers to regret minimization). For goal (a), we derive a problem-dependent sample complexity lower bound and discuss how to achieve matching algorithms. For goal (b), we provide a regret lower bound of $\Omega(T^{2/3})$ and provide a matching algorithm.

Efficient Change-Point Detection for Tackling Piecewise-Stationary Bandits [52]

We introduce GLR-klUCB, a novel algorithm for the piecewise i.i.d. 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 $O\big(\sqrt{TA\Upsilon_T \log(T)}\big)$ 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.

Adversarial Attacks on Linear Contextual Bandits [55]

Contextual bandit algorithms are applied in a wide range of domains, from advertising to recommender systems, from clinical trials to education. In many of these domains, malicious agents may have incentives to attack the bandit algorithm to induce it to perform a desired behavior. For instance, an unscrupulous ad publisher may try to increase their own revenue at the expense of the advertisers; a seller may want to increase the exposure of their products, or thwart a competitor's advertising campaign. In this paper, we study several attack scenarios and show that a malicious agent can force a linear contextual bandit algorithm to pull any desired arm $T - o(T)$ times over a horizon of T steps, while applying adversarial modifications to either rewards or contexts that only grow logarithmically as $O(\log T)$. We also investigate the case when a malicious agent is interested in affecting the behavior of the bandit algorithm in a single context (e.g., a specific user). We first provide sufficient conditions for the feasibility of the attack and we then propose an efficient algorithm to perform the attack. We validate our theoretical results on experiments performed on both synthetic and real-world datasets.

Forced-exploration free Strategies for Unimodal Bandits [58]

We consider a multi-armed bandit problem specified by a set of Gaussian or Bernoulli distributions endowed with a unimodal structure. Although this problem has been addressed in the literature [64], the state-of-the-art algorithms for such structure make appear a forced-exploration mechanism. We introduce IMED-UB, the first forced-exploration free strategy that exploits the unimodal-structure, by adapting to this setting the Indexed Minimum Empirical Divergence (IMED) strategy introduced by Honda and Takemura [66]. This strategy is proven optimal. We then derive KLUCB-UB, a KLUCB version of IMED-UB, which is also proven optimal. Owing to our proof technique, we are further able to provide a concise finite-time analysis of both strategies in an unified way. Numerical experiments show that both IMED-UB and KLUCB-UB perform similarly in practice and outperform the state-of-the-art algorithms.

Optimal Strategies for Graph-Structured Bandits [59]

We study a structured variant of the multi-armed bandit problem specified by a set of Bernoulli distributions $\nu = (\nu_{a,b})_{a\in\mathcal{A},b\in\mathcal{B}}$ with means $(\mu_{a,b})_{a\in\mathcal{A},b\in\mathcal{B}} \in [0,1]^{\mathcal{A}\times\mathcal{B}}$ and by a given weight matrix $\omega = (\omega_{b,b'})_{b,b'\in\mathcal{B}}$, where $\mathcal{A}$ is a finite set of arms and $\mathcal{B}$ is a finite set of users. The weight matrix ω is such that for any two users $b, b' \in \mathcal{B}$, $\max_{a\in\mathcal{A}} |\mu_{a,b} - \mu_{a,b'}| \leq \omega_{b,b'}$. This formulation is flexible enough to capture various situations, from highly-structured scenarios ($\omega \in \{0,1\}^{\mathcal{B}\times\mathcal{B}}$) to fully unstructured setups ($\omega \equiv 1$).We consider two scenarios depending on whether the learner chooses only the actions to sample rewards from or both users and actions. We first derive problem-dependent lower bounds on the regret for this generic graph-structure that involves a structure dependent linear programming problem. Second, we adapt to this setting the Indexed Minimum Empirical Divergence (IMED) algorithm introduced by Honda and Takemura (2015), and introduce the IMED-GS* algorithm. Interestingly, IMED-GS* does not require computing the solution of the linear programming problem more than about $\log(T)$ times after T steps, while being provably asymptotically optimal. Also, unlike existing bandit strategies designed for other popular structures, IMED-GS* does not resort to an explicit forced exploration scheme and only makes use of local counts of empirical events. We finally provide numerical illustration of our results that confirm the performance of IMED-GS*.

On Multi-Armed Bandit Designs for Dose-Finding Trials [51]

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.

8.2 Reinforcement learning

Only Relevant Information Matters: Filtering Out Noisy Samples to Boost RL [3]

In reinforcement learning, policy gradient algorithms optimize the policy directly and rely on sampling efficiently an environment. Nevertheless, while most sampling procedures are based on direct policy sampling, self-performance measures could be used to improve such sampling prior to each policy update. Following this line of thought, we introduce SAUNA, a method where non-informative transitions are rejected from the gradient update. The level of information is estimated according to the fraction of variance explained by the value function: a measure of the discrepancy between V and the empirical returns. In this work, we use this metric to select samples that are useful to learn from, and we demonstrate that this selection can significantly improve the performance of policy gradient methods. In this paper: (a) We define SAUNA's metric and introduce its method to filter transitions. (b) We conduct experiments on a set of benchmark continuous control problems. SAUNA significantly improves performance. (c) We investigate how SAUNA reliably selects samples with the most positive impact on learning and study its improvement on both performance and sample efficiency.

Inferential Induction: A Novel Framework for Bayesian Reinforcement Learning [34]

Bayesian Reinforcement Learning (BRL) offers a decision-theoretic solution to the reinforcement learning problem. While "model-based" BRL algorithms have focused either on maintaining a posterior distribution on models, BRL "model-free" methods try to estimate value function distributions but make strong implicit assumptions or approximations. We describe a novel Bayesian framework, inferential induction, for correctly inferring value function distributions from data, which leads to a new family of BRL algorithms. We design an algorithm, Bayesian Backwards Induction (BBI), with this framework. We experimentally demonstrate that BBI is competitive with the state of the art. However, its advantage relative to existing BRL model-free methods is not as great as we have expected, particularly when the additional computational burden is taken into account.

Tightening Exploration in Upper Confidence Reinforcement Learning [8]

The upper confidence reinforcement learning (UCRL2) algorithm introduced in [67] is a popular method to perform regret minimization in unknown discrete Markov Decision Processes under the average-reward criterion. Despite its nice and generic theoretical regret guarantees, this algorithm and its variants have remained until now mostly theoretical as numerical experiments in simple environments exhibit long burn-in phases before the learning takes place. In pursuit of practical efficiency, we present UCRL3, following the lines of UCRL2, but with two key modifications: First, it uses state-of-the-art time-uniform concentration inequalities to compute confidence sets on the reward and (component-wise) transition distributions for each state-action pair. Furthermore, to tighten exploration, it uses an adap-tive computation of the support of each transition distribution, which in turn enables us to revisit the extended value iteration procedure of UCRL2 to optimize over distributions with reduced support by disregarding low probability transitions, while still ensuring near-optimism. We demonstrate, through numerical experiments in standard environments, that reducing exploration this way yields a substantial numerical improvement compared to UCRL2 and its variants. On the theoretical side, these key modifications enable us to derive a regret bound for UCRL3 improving on UCRL2, that for the first time makes appear notions of local diameter and local effective support, thanks to variance-aware concentration bounds.

"I'm sorry Dave, I'm afraid I can't do that" Deep Q-Learning From Forbidden Actions [43]

The use of Reinforcement Learning (RL) is still restricted to simulation or to enhance human-operated systems through recommendations. Real-world environments (e.g. industrial robots or power grids) are generally designed with safety constraints in mind implemented in the shape of valid actions masks or contingency controllers. For example, the range of motion and the angles of the motors of a robot can be limited to physical boundaries. Violating constraints thus results in rejected actions or entering in a safe mode driven by an external controller, making RL agents incapable of learning from their mistakes. In this paper, we propose a simple modification of a state-of-the-art deep RL algorithm (DQN), enabling learning from forbidden actions. To do so, the standard Q-learning update is enhanced with an extra safety loss inspired by structured classification. We empirically show that it reduces the number of hit

constraints during the learning phase and accelerates convergence to near-optimal policies compared to using standard DQN. Experiments are done on a Visual Grid World Environment and Text-World domain.

A Machine of Few Words Interactive Speaker Recognition with Reinforcement Learning [44]

Speaker recognition is a well known and studied task in the speech processing domain. It has many applications, either for security or speaker adaptation of personal devices. In this paper, we present a new paradigm for automatic speaker recognition that we call Interactive Speaker Recognition (ISR). In this paradigm, the recognition system aims to incrementally build a representation of the speakers by requesting personalized utterances to be spoken in contrast to the standard text-dependent or text-independent schemes. To do so, we cast the speaker recognition task into a sequential decision-making problem that we solve with Reinforcement Learning. Using a standard dataset, we show that our method achieves excellent performance while using little speech signal amounts. This method could also be applied as an utterance selection mechanism for building speech synthesis systems.

HIGhER: Improving instruction following with Hindsight Generation for Experience Replay [22]

Language creates a compact representation of the world and allows the description of unlimited situations and objectives through compositionality. While these characterizations may foster instructing, conditioning or structuring interactive agent behavior, it remains an open-problem to correctly relate language understanding and reinforcement learning in even simple instruction following scenarios. This joint learning problem is alleviated through expert demonstrations, auxiliary losses, or neural inductive biases. In this paper, we propose an orthogonal approach called Hindsight Generation for Experience Replay (HIGhER) that extends the Hindsight Experience Replay approach to the language-conditioned policy setting. Whenever the agent does not fulfill its instruction, HIGhER learns to output a new directive that matches the agent trajectory, and it relabels the episode with a positive reward. To do so, HIGhER learns to map a state into an instruction by using past successful trajectories, which removes the need to have external expert interventions to relabel episodes as in vanilla HER. We show the efficiency of our approach in the BabyAI environment, and demonstrate how it complements other instruction following methods.

Fictitious Play for Mean Field Games: Continuous Time Analysis and Applications [57]

In this paper, we deepen the analysis of continuous time Fictitious Play learning algorithm to the consideration of various finite state Mean Field Game settings (finite horizon, γ-discounted), allowing in particular for the introduction of an additional common noise. We first present a theoretical convergence analysis of the continuous time Fictitious Play process and prove that the induced exploitability decreases at a rate $O(\frac{1}{t})$. Such analysis emphasizes the use of exploitability as a relevant metric for evaluating the convergence towards a Nash equilibrium in the context of Mean Field Games. These theoretical contributions are supported by numerical experiments provided in either model-based or model-free settings. We provide hereby for the first time converging learning dynamics for Mean Field Games in the presence of common noise.

Regret bounds for kernel-based reinforcement learning [53]

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 $O(H^3K^{max(12,\frac{2d}{2d+1})})$, where d is the covering dimension of the joint state-action space. We empirically validate Kernel-UCBVI on discrete and continuous MDPs.

CopyCAT: Taking Control of Neural Policies with Constant Attacks [32]

We propose a new perspective on adversarial attacks against deep reinforcement learning agents. Our main contribution is CopyCAT, a targeted attack able to consistently lure an agent into following an outsider's policy. It is pre-computed, therefore fast inferred, and could thus be usable in a real-time

scenario. We show its effectiveness on Atari 2600 games in the novel read-only setting. In this setting, the adversary cannot directly modify the agent's state – its representation of the environment – but can only attack the agent's observation – its perception of the environment. Directly modifying the agent's state would require a write-access to the agent's inner workings and we argue that this assumption is too strong in realistic settings.

Self-Attentional Credit Assignment for Transfer in Reinforcement Learning [25]

The ability to transfer knowledge to novel environments and tasks is a sensible desiderata for general learning agents. Despite the apparent promises, transfer in RL is still an open and little exploited research area. In this paper, we take a brand-new perspective about transfer: we suggest that the ability to assign credit unveils structural invariants in the tasks that can be transferred to make RL more sample efficient. Our main contribution is Secret, a novel approach to transfer learning for RL that uses a backward-view credit assignment mechanism based on a self-attentive architecture. Two aspects are key to its generality: it learns to assign credit as a separate offline supervised process and exclusively modifies the reward function. Consequently, it can be supplemented by transfer methods that do not modify the reward function and it can be plugged on top of any RL algorithm.

8.3 Adaptive control

Robust-Adaptive Control of Linear Systems: beyond Quadratic Costs [36]

We consider the problem of robust and adaptive model predictive control (MPC) of a linear system, with unknown parameters that are learned along the way (adaptive), in a critical setting where failures must be prevented (robust). This problem has been studied from different perspectives by different communities. However, the existing theory deals only with the case of quadratic costs (the LQ problem), which limits applications to stabilisation and tracking tasks only. In order to handle more general (non-convex) costs that naturally arise in many practical problems, we carefully select and bring together several tools from different communities, namely non-asymptotic linear regression, recent results in interval prediction, and tree-based planning. Combining and adapting the theoretical guarantees at each layer is non trivial, and we provide the first end-to-end suboptimality analysis for this setting. Interestingly, our analysis naturally adapts to handle many models and combines with a data-driven robust model selection strategy, which enables to relax the modelling assumptions. Last, we strive to preserve tractability at any stage of the method, that we illustrate on two challenging simulated environments.

Robust-Adaptive Interval Predictive Control for Linear Uncertain Systems [37]

We consider the problem of stabilization of a linear system, under state and control constraints, and subject to bounded disturbances and unknown parameters in the state matrix. First, using a simple least square solution and available noisy measurements, the set of admissible values for parameters is evaluated. Second, for the estimated set of parameter values and the corresponding linear interval model of the system, two interval predictors are recalled and an unconstrained stabilizing control is designed that uses the predicted intervals. Third, to guarantee the robust constraint satisfaction, a model predictive control algorithm is developed, which is based on solution of an optimization problem posed for the interval predictor. The conditions for recursive feasibility and asymptotic performance are established. Efficiency of the proposed control framework is illustrated by numeric simulations.

8.4 Applications

The challenge of controlling microgrids in the presence of rare events with Deep Reinforcement Learning [15]

The increased penetration of renewable energies and the need to decarbonize the grid come with a lot of challenges. Microgrids, power grids that can operate independently from the main system, are seen as a promising solution. They range from a small building to a neighbourhood or a village. As they co-locate generation, storage and consumption, microgrids are often built with renewable energies. At the same time, because they can be disconnected from the main grid, they can be more resilient and less dependent on central generation. Due to their diversity and distributed nature, advanced metering

and control will be necessary to maximize their potential. This paper presents a reinforcement learning algorithm to tackle the energy management of an off-grid microgrid, represented as a Markov Decision Process. The main objective function of the proposed algorithm is to minimize the global operating cost. By nature, rare events occur in physical systems. One of the main contribution of this paper is to demonstrate how to train agents in the presence of rare events. We prove that merging the combined experience replay method with a novel methods called "Memory Counter" unstucks the agent during its learning phase. Compared to baselines, we show that an extended version of Double Deep Q-Network with a priority list of actions into the decision making strategy process lowers significantly the operating cost. Experiments are conducted using two years of real-world data from Ecole Polytechnique in France.

Geometric Deep Reinforcement Learning for Dynamic DAG Scheduling [30]

In practice, it is quite common to face combinatorial optimization problems which contain uncertainty along with non-determinism and dynamicity. These three properties call for appropriate algorithms; reinforcement learning (RL) is dealing with them in a very natural way. Today, despite some efforts, most real-life combinatorial optimization problems remain out of the reach of reinforcement learning algorithms. In this paper, we propose a reinforcement learning approach to solve a realistic scheduling problem, and apply it to an algorithm commonly executed in the high performance computing community, the Cholesky factorization. On the contrary to static scheduling, where tasks are assigned to processors in a predetermined ordering before the beginning of the parallel execution, our method is dynamic: task allocations and their execution ordering are decided at runtime, based on the system state and unexpected events, which allows much more flexibility. To do so, our algorithm uses graph neural networks in combination with an actor-critic algorithm (A2C) to build an adaptive representation of the problem on the fly. We show that this approach is competitive with state-of-the-art heuristics used in high-performance computing runtime systems. Moreover, our algorithm does not require an explicit model of the environment, but we demonstrate that extra knowledge can easily be incorporated and improves performance. We also exhibit key properties provided by this RL approach, and study its transfer abilities to other instances.

Interdisciplinary Research in Artificial Intelligence: Challenges and Opportunities [14]

The use of artificial intelligence (AI) in a variety of research fields is speeding up multiple digital revolutions, from shifting paradigms in healthcare, precision medicine and wearable sensing, to public services and education offered to the masses around the world, to future cities made optimally efficient by autonomous driving. When a revolution happens, the consequences are not obvious straight away and, to date, there is no uniformly adapted framework to guide AI research to ensure a sustainable societal transition. To answer this need, here we analyze three key challenges to interdisciplinary AI research, and deliver three broad conclusions: 1) future development of AI should not only impact other scientific domains but should also take inspiration and benefit from other fields of science, 2) AI research must be accompanied by decision explainability, dataset bias transparency aswell as development of evaluation methodologies and creation of regulatory agencies to ensure responsibility, and 3) AI education should receive more attention, efforts and innovation from the educational and scientific communities. Our analysis is of interest not only to AI practitioners but also to other researchers and the general public as it offers ways to guide the emerging collaborations and interactions toward the most fruitful outcomes.

International electronic health record-derived COVID-19 clinical course profiles: the 4CE consortium [11]

We leveraged the largely untapped resource of electronic health record data to address critical clinical and epidemiological questions about Coronavirus Disease 2019 (COVID-19). To do this, we formed an international consortium (4CE) of 96 hospitals across 5 countries (https://www.covidclinical.net/). Contributors utilized the Informatics for Integrating Biology and the Bedside (i2b2) or Observational Medical Outcomes Partnership (OMOP) platforms to map to a common data model. The group focused on comorbidities and temporal changes in key laboratory test values. Harmonized data were analyzed locally and converted to a shared aggregate form for rapid analysis and visualization of regional differences and global commonalities. Data covered 27,584 COVID-19 cases with 187,802 laboratory tests. Case counts and laboratory trajectories were concordant with existing literature. Laboratory tests at the time of

diagnosis showed hospital-level differences equivalent to country-level variation across the consortium partners. Despite the limitations of decentralized data generation, we established a framework to capture the trajectory of COVID-19 disease in patients and their response to interventions.

Machine learning applications in drug development [16]

Due to the huge amount of biological and medical data available today, along with well-established machine learning algorithms, the design of largely automated drug development pipelines can now be envisioned. These pipelines may guide, or speed up, drug discovery; provide a better understanding of diseases and associated biological phenomena; help planning preclinical wet-lab experiments, and even future clinical trials. This automation of the drug development process might be key to the current issue of low productivity rate that pharmaceutical companies currently face. In this survey, we will particularly focus on two classes of methods: sequential learning and recommender systems, which are active biomedical fields of research.

8.5 Other

Restarted Bayesian Online Change-point Detector achieves Optimal Detection Delay [17]

In this paper, we consider the problem of sequential change-point detection where both the change-points and the distributions before and after the change are assumed to be unknown. For this problem of primary importance in statistical and sequential learning theory, we derive a variant of the Bayesian Online Change Point Detector proposed by [65] which is easier to analyze than the original version while keeping its powerful message-passing algorithm. We provide a non-asymptotic analysis of the false-alarm rate and the detection delay that matches the existing lower-bound. We further provide the first explicit high-probability control of the detection delay for such approach. Experiments on synthetic and realworld data show that this proposal outperforms the state-of-art change-point detection strategy, namely the Improved Generalized Likelihood Ratio (Improved GLR) while compares favorably with the original Bayesian Online Change Point Detection strategy.

8.6 Covid crisis

During the Inria-Covid, we helped the ScikitEDS team organize the data of Paris hospitals (AP-HP) under the form of daily dashboards. This was not only a matter of reporting but also helping practitioners (surgeons, biostatisticians, epidemiologists, and infectious disease specialists) run statistic models and propose new ML methods. This led to a couple of publications. We also led a work package in the EIT Health Covidom Community focused on the prediction of clinical worsening from symptoms on the telemonitoring app Covidom. With the help of Vianney Taquet and Clémence Léguillette who were hired on this project, we proposed new algorithms that are currently being implemented in production.

9 Bilateral contracts and grants with industry

9.1 Bilateral contracts with industry

- 2 contracts with Google regarding PhDs of J. Ferret and L. Hussenot (2020–2022), contract headed and PhD supervision by Ph. Preux.
- 1 contract with Facebook AI Research regarding PhD of J. Tarbouriech (2019–2021), contract headed and PhD supervision by Ph. Preux.
- 1 contract with Renault regarding PhD of É. Leurent (2018–2020), contract headed by Ph. Preux, PhD supervision by O-A. Maillard and D. Efimov (Valse, Inria Lille).
- N. Grinsztajn advised start-up Deeplife in Paris.
- J.-J. Vie is advising Axiome from Inria Startup Studio in since hackAtech Lille. With this startup, we competed to the NeurIPS 2020 Education Challenge and got ranked 5th over 23 submissions.

10 Partnerships and cooperations

10.1 International initiatives

10.1.1 Inria associate team not involved in an IIL

Associate team "Data Collection for Smart Crop Management" (DC4SCM) has begun in 2020. The partner being in India, the activities have heavily suffered from the covid-19 pandemic.

Scool also participates in the associate team 6PAC with CWI, headed by B. Guedj.

10.1.2 Inria international partners

Informal international partners

- with CGIAR, regarding agricultural practices recommendation.
- with A. Gopalan, IISC. Bangalore, about Markov decision processes with exponential family models.
- with Y. Bergner (New York University, USA), P. Halpin (University of North Carolina at Chapel Hill, USA), about multidimensional item response theory
- with A. Gilra and E. Vasilaki (University of Sheffield, United Kingdom), and with M. Große-Wenstrup (University of Vienna, Austria), we designed the proposal of the Chist-Era project CausalXRL, which has been accepted and begins on April 1st, 2021.
- with L. Martinez (Inria-Chile) *et al.*, we designed the proposal of the STIC AmSud project named EMISTRAL, which has been accepted and begins in 2021.
- with K. Meel and B. Ghosh (National University of Singapore), about formal fairness verification of machine learning algorithms.
- with I. Trummer (Cornell University, USA), about designing theory and algorithms of automated database optimizers based on reinforcement learning.
- with C. Dimitrakakis (University of Oslo, Norway), about designing fair and risk-sensitive reinforcement learning algorithms.
- with I.M Hassellöv (Chalmers University of Technology, Sweden), about using machine learning for enabling marine sciences and study of corresponding environmental phenomena.
- with Hisashi Kashima and Koh Takeuchi (Kyoto University, Japan), about active learning for education; this led to the proposal of a new associate team (OPALE° which has been accepted and starts in 20201.

10.2 International research visitors

10.2.1 Visits of international scientists

Prof. Anders Jonsson, University Pompeu Fabra (Spain), spent 1 year in the team on sabbatical, 2019–2020.

10.3 European initiatives

10.3.1 Collaborations in European programs, except FP7 and H2020

- Chist-Era Delta, headed by A. Jonsson (University Pompeu Fabra, Spain), local head: É. Kaufmann, 10/2017 – 12/2021.
- EIT Health Covidom Community, with AP-HP, headed by Patrick Jourdain, local head: J.-J. Vie, 4/2020 – 12/2020.

10.4 National initiatives

Scool is involved in 2 ANR projects:

- ANR Bold, headed by V. Perchet (ENS Paris-Saclay, ENSAE), local head: É. Kaufmann, 2019–2023.
- ANR JCJC Badass, O.-A. Maillard, 2016–2020.

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:

- T. Levent, PhD student, LPICM, École Polytechnique, control of smartgrids.
- R. Gautron, PhD student, Cirad, agricultural practices recommendation.
- É. Oyallon, CR CNRS, Sorbonne Université, machine learning on graphs.
- M. Valko, researcher DeepMind.
- K. Naudin, Aida Unit, Cirad Montpellier, agroecology.
- Y. Yordanov, A. Dechartres, A. Dinh, P. Jourdain, AP-HP, EIT Health Covidom Community
- A. Gramfort, G. Varoquaux, O. Grisel, Inria Saclay Parietal, projet Inria-Covid ScikitEDS avec AP-HP
- R. Khonsari, É. Vibert, A. Diallo, É. Vicaut, M. Bernaux, R. Nizard, T. Simon, N. Paris, L.-B. Luong, J. Assouad, C. Paugam, AP-HP, risks of mortality in surgery of COVID-19 patients
- P.-A. Jachiet, M. Doutreligne (DREES, then HAS), A. Floyrac (DREES, then Health Data Hub), synthetic data generation of the Système national de données de santé (SNDS)
- A. Delahaye-Duriez, INSERM, Université de Paris

10.5 Regional initiatives

- F. Pattou (PRU) and his group the Translational Research Laboratory for Diabetes (INSERM UMR 1190), CHU Lille, about patient personalized follow-up. This collaboration is funded by a set of projects:

  - project B4H from I-Site Lille,
  - project Phenomix from I-Site Lille,
  - project Perso-Surg funded by the CPER.

- A. Cotten CHU Lille and her group, project RAID, funded by the CPER.

- E. Chatelain, Eng. BiLille, medical data analysis.
- P. Schegg (PhD student), Ch. Duriez (DR Inria), J. Desquidt (MCF UdL), EPC Defrost Inria Lille, reinforcement learning for soft robotics in a surgical environment
- N. Mitton, DR Inria Lille, EPI Fun, data collection for smart crop management (associate team DC4SCM).
- D. Efimov, DR Inria Lille, EPC Valse, control theory (co-supervision of a PhD).
- project Biodimètre from Métropôle Européenne de Lille (MEL).

11 Dissemination

11.1 Promoting scientific activities

11.1.1 Scientific events: organisation

General chair, scientific chair

- J.-J. Vie: General Chair of EDM 2021

Member of the organizing committees J.J. Vie co-organized the following events:

- WASL 2020, Optimizing Human Learning, Third Workshop eliciting Adaptive Sequences for Learning, fully virtual, colocated with AIED 2020, with Benoît Choffin, Fabrice Popineau (LRI), Hisashi Kashima (Kyoto University, Japan), 6 July 2020, 12 participants
- FATED 2020, Fairness, Accountability, and Transparency in Educational Data, fully virtual, colocated with EDM 2020, with Nigel Bosch (University of Illinois at Urbana-Champaign, USA), Christopher Brooks (University of Michigan, USA), Shayan Doroudi (University of California Irvine, USA), Josh Gardner (University of Washington, USA), Kenneth Holstein (Carnegie Mellon University, USA), Andrew Lan (University of Massachusetts at Amherst, USA), Collin Lynch (North Carolina State University, USA), Beverly Park Woolf (University of Massachusetts at Amherst, USA), Mykola Pechenizkiy (Eindhoven University of Technology, The Netherlands), Steven Ritter (Carnegie Learning, USA), Renzhe Yu (University of California, Irvine, USA), 10 July 2020, 60 participants
- We organized two workshops related to e-learning: one about optimizing human learning at the AI for Education conference (AIED 2020), one about fairness in educational data mining at EDM 2020.

11.1.2 Scientific events: selection

Member of the conference program committees

- Ph. Preux: PC member for AAAI 2020, IJCAI 2020; Area chair for ECML (I declined being an area chair for ICML 2020)
- J.-J. Vie: Senior PC member of EDM 2020
- O-A. Maillard: PC member of COLT 2020, ICML 2020, ALT 2020.
- D. Basu: PC member of AAAI 2020.

Reviewer Scool members review paper submissions to all major conference in machine learning (e.g. ICML, NeurIPS, COLT, ALT, AI&Stats, ICLR, *etc*), AI (IJCAI, AAAI), and Privacy (PoPETS).

It should also be noted that due to the heavy load of work, we decline many invitations, even for these conferences.

11.1.3 Journal

Member of the editorial boards

- O-A. Maillard: part of the Journal of Machine Learning Research (JMLR) editorial board, as of July 2020.
- J.-J. Vie: part of the Journal of Educational Data Mining (JEDM) editorial board.

Reviewer - reviewing activities

- O-A. Maillard: Reviews for JMLR.
- J.-J. Vie: Reviews for JEDM.
- D. Basu: IEEE Access, IEEE Transactions on Dependable and Secure Computing.

11.1.4 Invited talks

Many events were canceled in 2020. Others could not be granted (e.g. O-A. Mailllard was invited for as a keynote speaker at SMMA 2020 but had to decline).

11.1.5 Scientific expertise

- O-A. Maillard:
  - member of the CRCN/ISFP jury at Inria Lille.
  - was asked reviewing expertise for the Astrid ANR committee.
  - is part of Commission Emploi Recherche (CER) in 2020.
- Ph. Preux:
  - member of the national Inria DR jury.
  - member of the CRCN/ISFP jury at Inria Grenoble.
  - is a member of the scientific committee “data science and models” (CSS5) of the IRD.
  - is a member of a group thinking about “responsible AI” in REV3, in Lille.
  - He also declined reviewing many requests (including for the ANR, and IUF).

11.1.6 Research administration

- Philippe Preux is « Délégué Scientifique Adjoint » of Inria-Lille.

  As such, he is a member of the « Commission d’évaluation » of Inria.

  As head of Scool (REP), he is a member of the « Comité des équipes-projets » (CEP)of Inria-Lille.

  He is also a member of the « Bureau scientifique du centre » (BSC) of Inria-Lille.

11.2 Teaching - Supervision - Juries

11.2.1 Teaching

- D. Basu: Learning to Optimise Online with Full and Partial Information, Reading Session, M2 Data Science, UdL.
- D. Baudry taught about 100 h. in 2020, in data science and NLP, in M1 maths and M2 web analyst at the UdL.
- O. Darwiche taught reinforcement learning at École Centrale de Lille and in the data science master of the Université de Lille. He also served as a TA in reinforcement learning at the African Institute for Mathematical Sciences (AIMS) in Ghana in February.

- É. Kaufmann: Data Mining (21h) M1 Maths/Finance, UdL.
- É. Kaufmann: Sequential Decision Making (24h) M2 Data Science, Centrale Lille
- O-A. Maillard: Statistical Reinforcement Learning (42h), MAP/INF641, Master Artificial Intelligence and advanced Visual Computing, École Polytechnique.
- R. Ouhamma taught about 40 hours, algorithmics and programming at UdL, and also proba/stats in M2 at Centrale-Lille.
- S. Perrin taught about 30 hours, Unix and databases at UdL.
- Ph. Preux: « IA et apprentissage automatique », DU IA & Santé, UdL
- H. Saber: as part of his agrégé de mathématiques duty, he taught in the licence and master of mathematics at UdL.
- M. Seurin is ATER, hence teaches 192 hours during the academic year at UdL. He taught machine learning, data science, reinforcement learning and other topics in licence and master MIASHS.
- F. Valentini: « deep learning pour la radiologie », practical session, DU IA & Santé, UdL.
- J.-J. Vie: Introduction to Machine Learning (24h), M2 Mechanical Engineering, Polytech'Lille.
- J.-J. Vie: Deep Learning Do It Yourself (45h), M1, ENS Paris.

Due to the Covid-19 crisis, our participation to some summer schools has been canceled.

11.2.2 Supervision

Apart from Ph.D. students, we supervised the following students in 2020:

- Ph. Preux supervised A. Zeddoun, A. Moulin (master interns), H. Delavenne, A. Tuynmann, A. Vigneron (L3 interns).
- J.-J. Vie supervised Pierre Bourse (L3), Aymeric Floyrac, Salim Nadir (M2).

In Scool, we consider that supervising students is part of the training of a Ph.D. student. Therefore, some of them (like Y. Flet-Berliac and R. Ouhamma) have participated to supervision of master students, under the supervision of a permanent researcher.

11.2.3 Juries

- É. Kaufmann was part of the following juries:
  - PhD: Erwan LeCarpentier, IRIT Toulouse, July
  - PhD: Yann Issartel, Université Paris-Saclay, December
  - PhD: Margaux Brégère, Université Paris-Saclay, December
  - PhD: Andrea Tirinzoni, Polytechnico Milano, Italy, reviewer
- Ph. Preux was part of the following juries:
  - hdr: Nistor Grozavu, Université Paris 13, March, reviewer
  - hdr: Sylvain Lamprier, Sorbonne Université, September, reviewer
  - hdr: Émilie Kaufmann, UdL, November, president
  - PhD: Tanguy Levent, École Polytechnique, December, president
  - PhD: Matthieu Jedor, Université Paris-Saclay, December, reviewer
- O-A. Maillard was part of the following juries:
  - PhD: Robin Vogel, Télécom-Paris, October, examiner
  - PhD: Margaux Brégère, Université Paris-Saclay, December, reviewer.

11.3 Popularization

- O-A. Maillard, communication on « Jardinage massivement collaboratif » related to his Action Exploratoire SR4SG during French fête de la science week, October 2020. https://www.youtube.com/watch?v=AJHA1kG2d9A
- É. Leurent, communication at “Inria 13:45” event, October 13.
- Ph. Preux is a member of the Collectif des chercheurs *« œuvres et recherches »* of UdL, in particular the AI working group. The goal is to investigate links between AI and artists, organize events etc.

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] G. A. Brat, G. M. Weber, N. Gehlenborg, P. Avillach, N. P. Palmer, L. Chiovato, J. Cimino, B. K. Beaulieu-Jones, S. L'Yi, M. S. Keller et al. 'International electronic health record-derived COVID-19 clinical course profiles: the 4CE consortium'. In: *npj Digital Medicine* 3.1 (Dec. 2020), #109. DOI: 10.1038/s41746-020-00308-0. URL: https://hal.archives-ouvertes.fr/hal-02918344.

[12] 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.

[13] T. Kocák, R. Munos, B. Kveton, S. Agrawal and M. Valko. 'Spectral bandits'. In: *Journal of Machine Learning Research* (2020). URL: https://hal.inria.fr/hal-03084249.

[14] R. Kusters, D. Misevic, H. Berry, A. Cully, Y. Le Cunff, L. Dandoy, N. Díaz-Rodríguez, M. Ficher, J. Grizou, A. Othmani, T. Palpanas, M. Komorowski, P. Loiseau, C. Moulin-Frier, S. Nanini, D. Quercia, M. Sebag, F. Soulié Fogelman, S. Taleb, L. Tupikina, V. Sahu, J.-J. Vie and F. Wehbi. 'Interdisciplinary Research in Artificial Intelligence: Challenges and Opportunities'. In: *Frontiers in Big Data* 3 (23rd Nov. 2020). DOI: 10.3389/fdata.2020.577974. URL: https://hal.inria.fr/hal-03111148.

[15] T. Levent, P. Preux, G. Henri, R. Alami, P. Cordier and Y. Bonnassieux. 'The challenge of controlling microgrids in the presence of rare events with Deep Reinforcement Learning'. In: *IET Smart Grid* (2020). DOI: 10.1049/stg2.12003. URL: https://hal.archives-ouvertes.fr/hal-02971554.

[16] C. Réda, E. Kaufmann and A. Delahaye-Duriez. 'Machine learning applications in drug development'. In: *Computational and Structural Biotechnology Journal* 18 (2020), pp. 241–252. DOI: 10.1016/j.csbj.2019.12.006. URL: https://hal.archives-ouvertes.fr/hal-02533303.

International peer-reviewed conferences

[17] R. Alami, O.-A. Maillard and R. Féraud. 'Restarted Bayesian Online Change-point Detector achieves Optimal Detection Delay'. In: International Conference on Machine Learning. Wien, Austria, July 2020. URL: https://hal.archives-ouvertes.fr/hal-03021712.

[18] 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.

[19] D. Baudry, E. Kaufmann and O.-A. Maillard. 'Sub-sampling for Efficient Non-Parametric Bandit Exploration'. In: NeurIPS 2020. Vancouver, Canada, 7th Dec. 2020. URL: https://hal.archives-ouvertes.fr/hal-02977552.

[20] E. Boursier, E. Kaufmann, A. Mehrabian and V. Perchet. 'A Practical Algorithm for Multiplayer Bandits when Arm Means Vary Among Players'. In: AISTATS 2020 - 23rd International Conference on Artificial Intelligence and Statistics. Palermo, Italy, 26th Aug. 2020. URL: https://hal.archives-ouvertes.fr/hal-02006069.

[21] J. Cheshire, P. Ménard and A. Carpentier. 'The Influence of Shape Constraints on the Thresholding Bandit Problem'. In: COLT 2020 - Thirty Third Conference on Learning Theory. Vol. 125. Graz / Virtual, Austria, 2020, pp. 1228–1275. URL: https://hal.archives-ouvertes.fr/hal-03001947.

[22] G. Cideron, M. Seurin, F. Strub and O. Pietquin. 'HIGhER: Improving instruction following with Hindsight Generation for Experience Replay'. In: ADPRL 2020 - IEEE SSCI Conference on Adaptive Dynamic Programming and Reinforcement Learning. Camberra / Virtual, Australia, 1st Dec. 2020. URL: https://hal.archives-ouvertes.fr/hal-03123981.

[23] 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.

[24] R. Degenne, P. Ménard, X. Shang and M. Valko. 'Gamification of pure exploration for linear bandits'. In: International Conference on Machine Learning. Vienna / Virtual, Austria, 2020. URL: https://hal.archives-ouvertes.fr/hal-02884330.

[25] J. Ferret, R. Marinier, M. Geist and O. Pietquin. 'Self-Attentional Credit Assignment for Transfer in Reinforcement Learning'. In: IJCAI 2020 - 29th International Joint Conference on Artificial Intelligence. Yokohama / Virtual, Japan, 11th July 2020. URL: https://hal.inria.fr/hal-03159832.

[26] 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.

[27] 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.

[28] 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.

[29] 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, 11th July 2020. DOI: 10.24963/ijcai.2020/376. URL: https://hal.inria.fr/hal-02091547.

[30] N. Grinsztajn, O. Beaumont, E. Jeannot and P. Preux. 'Geometric Deep Reinforcement Learning for Dynamic DAG Scheduling'. In: IEEE SSCI 2020 - Symposium Series on Computational Intelligence. SSCI 2020 proceedings. Canberra / Virtual, Australia, Dec. 2020. URL: https://hal.inria.fr/hal-03028981.

[31] 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.

[32] L. Hussenot, M. Geist and O. Pietquin. 'CopyCAT: Taking Control of Neural Policies with Constant Attacks'. In: AAMAS 2020 - 19th International Conference on Autonomous Agents and Multi-Agent Systems. Virtual, New Zealand, 9th May 2020. URL: https://hal.inria.fr/hal-03162124.

[33] A. Jonsson, E. Kaufmann, P. Ménard, O. D. Domingues, E. Leurent and M. Valko. 'Planning in Markov Decision Processes with Gap-Dependent Sample Complexity'. In: Neural Information Processing Systems. Vancouver, France, 7th Dec. 2020. URL: https://hal.archives-ouvertes.fr/hal-02863486.

[34] E. Jorge, H. Eriksson, C. Dimitrakakis, D. Basu and D. Grover. 'Inferential Induction: A Novel Framework for Bayesian Reinforcement Learning'. In: *"I Can't Believe It's Not Better!" at NeurIPS Workshops*. "I Can't Believe It's Not Better!" at NeurIPS Workshops. Vol. 137. Proceedings of Machine Learning Research. Vancouver, Canada, 12th Dec. 2020, pp. 43–52. URL: https://hal.archives-ouvertes.fr/hal-03125100.

[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] E. Leurent, D. Efimov and O.-A. Maillard. 'Robust-Adaptive Control of Linear Systems: beyond Quadratic Costs'. In: NeurIPS 2020 - 34th Conference on Neural Information Processing Systems. Vancouver / Virtual, Canada, 6th Dec. 2020. URL: https://hal.inria.fr/hal-03004060.

[37] E. Leurent, D. Efimov and O.-A. Maillard. 'Robust-Adaptive Interval Predictive Control for Linear Uncertain Systems'. In: CDC 2020 - 59th IEEE Conference on Decision and Control. Jeju Island / Virtual, South Korea, 10th Dec. 2020. URL: https://hal.inria.fr/hal-02942414.

[38] E. Leurent and O.-A. Maillard. 'Monte-Carlo Graph Search: the Value of Merging Similar States'. In: ACML 2020 - 12th Asian Conference on Machine Learning. Vol. 129. Bangkok / Virtual, Thailand, 2020, pp. 577–602. URL: https://hal.inria.fr/hal-03004124.

[39] 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.

[40] P. Perrault, E. Boursier, V. Perchet and M. Valko. 'Statistical efficiency of Thompson sampling for combinatorial semi-bandits'. In: Neural Information Processing Systems. virtual, France, 2020. URL: https://hal.archives-ouvertes.fr/hal-03089794.

[41] P. Perrault, J. Healey, Z. Wen and M. Valko. 'Budgeted online influence maximization'. In: International Conference on Machine Learning. Vienna, Austria, 2020. URL: https://hal.archives-ouvertes.fr/hal-02904278.

[42] P. Perrault, V. Perchet and M. Valko. 'Covariance-adapting algorithm for semi-bandits with application to sparse outcomes'. In: Conference on Learning Theory. Graz, Austria, 2020. URL: https://hal.archives-ouvertes.fr/hal-02876102.

[43] M. Seurin, P. Preux and O. Pietquin. '"I'm sorry Dave, I'm afraid I can't do that" Deep Q-Learning From Forbidden Actions'. In: Internationnal Joint Conference on Neural Networks. Glasgow, United Kingdom, 17th July 2020. URL: https://hal.inria.fr/hal-02387419.

[44] M. Seurin, F. Strub, P. Preux and O. Pietquin. 'A Machine of Few Words Interactive Speaker Recognition with Reinforcement Learning'. In: *Interspeech 2020 proceedings*. Conference of the International Speech Communication Association (INTERSPEECH). Shanghai, China, 25th Oct. 2020. DOI: 10.21437/Interspeech.2020-2892. URL: https://hal.archives-ouvertes.fr/hal-03123999.

[45] X. Shang, R. De Heide, E. Kaufmann, P. Ménard and M. Valko. 'Fixed-confidence guarantees for Bayesian best-arm identification'. In: International Conference on Artificial Intelligence and Statistics. Palermo, Italy, 2020. URL: https://hal.archives-ouvertes.fr/hal-02330187.

[46] C. Trinh, E. Kaufmann, C. Vernade and R. Combes. 'Solving Bernoulli Rank-One Bandits with Unimodal Thompson Sampling'. In: ALT 2020 - 31st International Conference on Algorithmic Learning Theory. Vol. 117. San Diego, United States, 8th Feb. 2020, pp. 1–28. URL: https://hal.archives-ouvertes.fr/hal-02396943.

Conferences without proceedings

[47] 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.

Doctoral dissertations and habilitation theses

[48] E. Leurent. 'Safe and Efficient Reinforcement Learning for Behavioural Planning in Autonomous Driving'. Université de Lille, 30th Oct. 2020. URL: https://hal.inria.fr/tel-03035705.

[49] P. Perrault. 'Efficient Learning in Stochastic Combinatorial Semi-Bandits'. Univeristé Paris-Saclay, 30th Nov. 2020. URL: https://tel.archives-ouvertes.fr/tel-03093268.

[50] F. Strub. 'Multimodal and Interactive Models for Visually Grounded Language Learning'. Université de Lille; École doctorale, ED SPI 074 : Sciences pour l'Ingénieur, 28th Jan. 2020. URL: https://tel.archives-ouvertes.fr/tel-03018038.

Reports & preprints

[51] M. Aziz, E. Kaufmann and M.-K. Riviere. *On Multi-Armed Bandit Designs for Dose-Finding Trials*. 6th Apr. 2020. URL: https://hal.archives-ouvertes.fr/hal-02533297.

[52] L. Besson, E. Kaufmann, O.-A. Maillard and J. Seznec. *Efficient Change-Point Detection for Tackling Piecewise-Stationary Bandits*. 8th Dec. 2020. URL: https://hal.inria.fr/hal-02006471.

[53] O. D. Domingues, P. Ménard, M. Pirotta, E. Kaufmann and M. Valko. *Regret bounds for kernel-based reinforcement learning*. Vienna, Austria, 14th Apr. 2020. URL: https://hal.inria.fr/hal-02541790.

[54] 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.

[55] E. Garcelon, B. Roziere, L. Meunier, J. Tarbouriech, O. Teytaud, A. Lazaric and M. Pirotta. *Adversarial Attacks on Linear Contextual Bandits*. 27th Oct. 2020. URL: https://hal.archives-ouvertes.fr/hal-02979184.

[56] P. Ménard, O. D. Domingues, A. Jonsson, E. Kaufmann, E. Leurent and M. Valko. *Fast active learning for pure exploration in reinforcement learning*. DeepMind, 26th July 2020. URL: https://hal.inria.fr/hal-02906985.

[57] S. Perrin, J. Pérolat, M. Laurière, M. Geist, R. Elie and O. Pietquin. *Fictitious Play for Mean Field Games: Continuous Time Analysis and Applications*. 7th Sept. 2020. URL: https://hal.inria.fr/hal-02931977.

[58] H. Saber, P. Ménard and O.-A. Maillard. *Forced-exploration free Strategies for Unimodal Bandits*. 29th June 2020. URL: https://hal.archives-ouvertes.fr/hal-02883907.

[59] H. Saber, P. Ménard and O.-A. Maillard. *Optimal Strategies for Graph-Structured Bandits*. 8th July 2020. URL: https://hal.archives-ouvertes.fr/hal-02891139.

[60] X. Shang, H. Shao and J. Qian. *Stochastic bandits with vector losses: Minimizing ℓ^∞-norm of relative losses*. 15th Oct. 2020. URL: https://hal.archives-ouvertes.fr/hal-02968536.

12.3 Cited publications

[61] A. Baransi, O.-A. Maillard and S. Mannor. 'Sub-sampling for Multi-armed Bandits'. In: *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2014, Nancy, France, September 15–19, 2014. Proceedings, Part I*. Ed. by T. Calders, F. Esposito, E. Hüllermeier and R. Meo. Vol. 8724. Lecture Notes in Computer Science. Springer, 2014, pp. 115–131. DOI: 10.1007/978-3-662-44848-9_8. URL: https://doi.org/10.1007/978-3-662-44848-9_8.

[62] I. Bistritz and A. Leshem. 'Distributed Multi-Player Bandits — a Game of Thrones Approach'. In: *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3–8, 2018, Montréal, Canada*. Ed. by S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi and R. Garnett. 2018, pp. 7222–7232. URL: https://proceedings.neurips.cc/paper/2018/hash/c2964caac096f26db222cb325aa267cb-Abstract.html.

[63] H. P. Chan. 'The multi-armed bandit problem: An efficient nonparametric solution'. In: *The Annals of Statistics* 48.1 (2020), pp. 346–373. DOI: 10.1214/19-AOS1809. URL: https://doi.org/10.1214/19-AOS1809.

[64] R. Combes and A. Proutière. 'Unimodal Bandits: Regret Lower Bounds and Optimal Algorithms'. In: *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21–26 June 2014*. Vol. 32. JMLR Workshop and Conference Proceedings. JMLR.org, 2014, pp. 521–529. URL: http://proceedings.mlr.press/v32/combes14.html.

[65] P. Fearnhead and Z. Liu. 'On-Line Inference for Multiple Changepoint Problems'. In: *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 69.4 (2007), pp. 589–605. URL: http://www.jstor.org/stable/4623285.

[66] J. Honda and A. Takemura. 'Non-asymptotic analysis of a new bandit algorithm for semi-bounded rewards'. In: *J. Mach. Learn. Res.* 16 (2015), pp. 3721–3756. URL: http://dl.acm.org/citation.cfm?id=2912115.

[67] T. Jaksch, R. Ortner and P. Auer. 'Near-optimal Regret Bounds for Reinforcement Learning'. In: *J. Mach. Learn. Res.* 11 (2010), pp. 1563–1600. URL: http://portal.acm.org/citation.cfm?id=1859902.

[68] S. Katariya, B. Kveton, C. Szepesvári, C. Vernade and Z. Wen. 'Bernoulli Rank-1 Bandits for Click Feedback'. In: *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19–25, 2017*. Ed. by C. Sierra. ijcai.org, 2017, pp. 2001–2007. DOI: 10.24963/ijcai.2017/278. URL: https://doi.org/10.24963/ijcai.2017/278.

[69] S. Katariya, B. Kveton, C. Szepesvári, C. Vernade and Z. Wen. 'Stochastic Rank-1 Bandits'. In: *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017, 20–22 April 2017, Fort Lauderdale, FL, USA*. Ed. by A. Singh and X. (Zhu. Vol. 54. Proceedings of Machine Learning Research. PMLR, 2017, pp. 392–401. URL: http://proceedings.mlr.press/v54/katariya17a.html.

[70] S. Paladino, F. Trovò, M. Restelli and N. Gatti. 'Unimodal Thompson Sampling for Graph-Structured Arms'. In: *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4–9, 2017, San Francisco, California, USA*. Ed. by S. P. Singh and S. Markovitch. AAAI Press, 2017, pp. 2457–2463. URL: http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14325.

[71] D. Russo. 'Simple Bayesian Algorithms for Best Arm Identification'. In: *Proceedings of the 29th Conference on Learning Theory, COLT 2016, New York, USA, June 23–26, 2016*. Ed. by V. Feldman, A. Rakhlin and O. Shamir. Vol. 49. JMLR Workshop and Conference Proceedings. JMLR.org, 2016, pp. 1417–1418. URL: http://proceedings.mlr.press/v49/russo16.html.