RESEARCH CENTRE

**Inria Saclay Centre at Université Paris-Saclay**

**IN PARTNERSHIP WITH:**

**CNRS, Université Paris-Saclay**

# 2024 ACTIVITY REPORT

# Project-Team CELESTE

## mathematical statistics and learning

**IN COLLABORATION WITH: Laboratoire de mathématiques d'Orsay de l'Université de Paris-Sud (LMO)**

**DOMAIN**

**Applied Mathematics, Computation and Simulation**

**THEME**

**Optimization, machine learning and statistical methods**

Inria

# Contents

# Project-Team CELESTE

*Creation of the Project-Team: 2019 June 01*

## Keywords

**Computer sciences and digital sciences**

A3.1.1. – Modeling, representation
A3.1.8. – Big data (production, storage, transfer)
A3.3. – Data and knowledge analysis
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
A3.4.7. – Kernel methods
A3.4.8. – Deep learning
A3.5.1. – Analysis of large graphs
A6.1. – Methods in mathematical modeling
A9.2. – Machine learning

**Other research topics and application domains**

B1.1.4. – Genetics and genomics
B1.1.7. – Bioinformatics
B2.2.4. – Infectious diseases, Virology
B2.3. – Epidemiology
B4. – Energy
B4.4. – Energy delivery
B4.5. – Energy consumption
B5.2.1. – Road vehicles
B5.2.2. – Railway
B5.5. – Materials
B5.9. – Industrial maintenance
B7.1. – Traffic management
B7.1.1. – Pedestrian traffic and crowds
B9.5.2. – Mathematics
B9.8. – Reproducibility
B9.9. – Ethics

# 1 Team members, visitors, external collaborators

## Research Scientists

- Kevin Bleakley [INRIA, Researcher]
- Etienne Boursier [INRIA, ISFP]
- Gilles Celeux [INRIA, Emeritus]
- Evgenii Chzhen [CNRS, Researcher]
- Gilles Stoltz [CNRS, Senior Researcher]

## Faculty Members

- Sylvain Arlot [Team leader, UNIV PARIS SACLAY, Professor]
- Christophe Giraud [UNIV PARIS SACLAY, Professor]
- Alexandre Janon [UNIV PARIS SACLAY, Associate Professor]
- Christine Keribin [UNIV PARIS SACLAY, Professor]
- Pascal Massart [UNIV PARIS SACLAY, Professor]
- Patrick Pamphile [UNIV PARIS SACLAY, Associate Professor]
- Marie-Anne Poursat [UNIV PARIS SACLAY, Associate Professor]
- Vincent Rivoirard [LMO, Professor Delegation, from Sep 2024]

## Post-Doctoral Fellow

- Pierre Humbert [UNIV PARIS SACLAY, until Jun 2024]

## PhD Students

- Aymeric Capitaine [UNIV PARIS SACLAY]
- Samy Clementz [SORBONNE UNIVERSITE]
- Bertrand Even [UNIV PARIS SACLAY, from Sep 2024]
- Guillermo Martin [UNIV PARIS SACLAY, from Oct 2024]
- Leonardo Martins Bianco [UNIV PARIS SACLAY]
- Chiara Mignacco [UNIV PARIS SACLAY]
- Pierre-Andre Mikem [UNIV PARIS SACLAY & METAFORA]
- Dhia Elhaq Ouerfelli [UNIV PARIS SACLAY, from Oct 2024]
- Guillaume Principato [UNIV PARIS SACLAY & EDF]
- Gayane Taturyan [IRT SYSTEM X]
- Daniil Tipakin [UNIV PARIS SACLAY & ECOLE POLYTECHNIQUE]
- Victor Turmel [UNIV PARIS SACLAY, from Sep 2024]

### Interns and Apprentices

- Dhia Elhaq Ouerfelli [INRIA, Intern, from Apr 2024 until Sep 2024]
- Shuailong Zhu [INRIA, Intern, from Oct 2024]

### Administrative Assistant

- Aissatou-Sadio Diallo [INRIA]

### External Collaborators

- Claire Lacour [UNIV PARIS EST]
- Jean-Michel Poggi [UNIV PARIS SACLAY]

## 2 Overall objectives

### 2.1 Mathematical statistics and learning

Data science—a vast field that includes statistics, machine learning, signal processing, data visualization, and databases—has become front-page news due to its ever-increasing impact on society, over and above the important role it already played in science over the last few decades. Within data science, the statistical community has long-term experience in how to infer knowledge from data, based on solid mathematical foundations. The recent field of machine learning has also made important progress by combining statistics and optimization, with a fresh point of view that originates in applications where prediction is more important than building models.

The Celeste project-team is positioned at the interface between statistics and machine learning. We are statisticians in a mathematics department, with strong mathematical backgrounds, interested in interactions between theory, algorithms, and applications. Indeed, applications are the source of many of our interesting theoretical problems, while the theory we develop plays a key role in (i) understanding how and why successful statistical learning algorithms work—hence improving them—and (ii) building new algorithms upon mathematical statistics-based foundations. Therefore, we tackle several major challenges of machine learning with our mathematical statistics point of view (in particular the algorithmic fairness issue), always having in mind that modern datasets are often high-dimensional and/or large-scale, which must be taken into account at the building stage of statistical learning algorithms. For instance, there often are trade-offs between statistical accuracy and complexity which we want to clarify as much as possible.

In addition, most theoretical guarantees that we prove are non-asymptotic, which is important because the number of features $p$ is often larger than the sample size $n$ in modern datasets, hence asymptotic results with $p$ fixed and $n \to +\infty$ are not relevant. The non-asymptotic approach is also closer to the real-world than specific asymptotic settings, since it is difficult to say whether $p = 1000$ and $n = 100$ corresponds to the setting $p = 10n$ or $p = n^{3/2}$.

Finally, a key ingredient in our research program is connecting our theoretical and methodological results with (a great number of) real-world applications. This is the reason why a large part of our work is devoted to industrial and medical data modeling on a set of real-world problems coming from our long-term collaborations with several partners, as well as various opportunistic one-shot collaborations.

## 3 Research program

### 3.1 General presentation

We split our research program into four research axes, distinguishing problems and methods that are traditionally considered part of mathematical statistics (e.g., model selection and hypothesis testing, see section 3.2) from those usually tackled by the machine learning community (e.g., multi-armed bandits

and reinforcement learning, deep learning, clustering and pairwise-data inference, see section 3.3). Section 3.4 is devoted to industrial and medical data modeling questions which arise from several long-term collaborations and more recent research contracts. Finally, section 3.5 is devoted to algorithmic fairness, a theme of Celeste which we want to specifically emphasize. Despite presenting mathematical statistics, machine learning, and data modeling as separate axes, we would like to make clear that these axes are strongly interdependent in our research and that this dependence is a key factor in our success.

## 3.2 Mathematical statistics

One of our main goals is to address major challenges in machine learning in which mathematical statistics naturally play a key role, in particular in the following two areas of research.

**Estimator selection.** Any machine learning procedure requires a choice for the values of hyper-parameters, and one must also choose among the numerous procedures available for any given learning problem; both situations correspond to an estimator selection problem. High-dimensional variable (feature) selection is another key estimator selection problem. Celeste addresses all such estimator selection problems, where the goal is to select an estimator (or a set of features) minimizing the prediction/estimation risk, and the corresponding non-asymptotic theoretical guarantee—which we want to prove in various settings—is an oracle inequality.

**Statistical reproducibility.** Science currently faces a reproducibility crisis, making it necessary to provide statistical inference tools (hypotheses tests, confidence regions) for assessing the significance of the output of any learning algorithm in a computationally efficient way. Our goal here is to develop methods for which we can prove upper bounds on the type I error rate, while maximizing the detection power under this constraint. We are particularly interested in the variable selection case, which here leads to a multiple testing problem for which key metrics are the family-wise error rate (FWER) and the false discovery rate (FDR).

## 3.3 Theoretical foundations of machine learning

Our distinguishing approach (compared to peer groups around the world) is to offer a statistical and mathematical point of view on machine-learning (ML) problems and algorithms. Our main focus is to provide theoretical guarantees for certain ML problems, with special attention paid to the statistical point of view, in particular minimax optimality and statistical adaptivity. In the areas of deep learning and big data, computationally-efficient optimization algorithms are essential. The choice of the optimization algorithm has been shown to have a dramatic impact on generalization properties of predictors. Such empirical observations have led us to investigate the interplay between computational efficiency and statistical properties. The set of problems we tackle includes online learning (expert aggregation, stochastic bandits, reinforcement learning), clustering and co-clustering, pairwise-data inference, semi-supervised learning, and the interplay between optimization and statistical properties.

## 3.4 Industrial and medical data modeling

Celeste collaborates with industry and with medicine/public health institutes to develop methods and apply results of a broadly statistical nature—whether they be prediction, aggregation, anomaly detection, forecasting, and so on—in relationship with pressing industrial and/or societal needs (see sections 4 and 5.2). Most of these methods and applied results are directly related to the more theoretical subjects examined in the first two research axes, including for instance estimator selection, aggregation, and supervised and unsupervised classification. Furthermore, Celeste is positioned well for problems with data requiring unconventional methods—for instance, non asymptotic analysis and data with selection bias—, and in particular problems that can give rise to technology transfers in the context of Cifre Ph.D.s.

### 3.5 Algorithmic fairness

Machine-learning algorithms make pivotal decisions which influence our lives on a daily basis, using data about individuals. Recent studies show that imprudent use of these algorithms may lead to unfair and discriminatory decisions, often inheriting or even amplifying disparities present in data. The goal of Celeste on this topic is to design and analyze novel tractable algorithms that, while still optimizing prediction performance, mitigate or remove unfair decisions of the learned predictor. A major challenge in the machine-learning fairness literature is to obtain algorithms which satisfy fairness and risk guarantees simultaneously. Several empirical studies suggest that there is a trade-off between the fairness and accuracy of a learned model: more accurate models are less fair. We are focused on providing user-friendly statistical quantification of such trade-offs and building statistically-optimal algorithms in this context, with special attention paid to the online learning setting. Relying on the strong mathematical and statistical competency of the team, we approach the problem from an angle that differs from the mainstream computer science literature.

## 4 Application domains

### 4.1 Electricity load consumption: forecasting and control

Celeste has a long-term collaboration with EDF R&D on electricity consumption. An important problem is to forecast consumption, e.g., for electric vehicles (EVs). We currently work on hierarchical consumption data of EVs, for which we aim to output probabilistic forecasts, e.g., through conformal inference methods.

### 4.2 Reliability

Data collected on the lifetime of complex systems is often non-homogeneous, affected by variability in component production and differences in real-world system use. In general, this variability is neither controlled nor observed in any way, but must be taken into account in reliability analysis. We use latent structure models to identify the main causes of failure, and to predict system reliability as accurately as possible [10].

### 4.3 Neglected tropical diseases

Celeste collaborates with researchers at Institut Pasteur on encephalitis in South-East Asia, especially with Jean-David Pommier.

### 4.4 Explainability in change-points detection in high dimensional multivariate time series

Detecting changes in time series is essential in many areas, such as identifying anomalies in industrial processes, monitoring medical conditions, detecting variations in climatic conditions, or analyzing fluctuations in financial markets.

Numerous change-point detection approaches have been developed, both offline and online, and applied to univariate and multivariate series. In the multivariate context, where the components of the series can represent the measurements of thousands of sensors, an important question remains after the change-point has been estimated: which sensors are specifically involved in the detected change?

In Dhia El Haq Ouerfelli's PhD thesis, we develop post-hoc methods to identify the coordinates involved in a detected change and to evaluate the quality of this detection.

### 4.5 Cytometry

Celeste collaborates with Metafora to explore the use of multiple instance learning in flow cytometry as a means of early detection of specific cancers. This is in collaboration with Pascal Massart and Christine Keribin, in the context of Pierre-André Mikem's Cifre PhD, which follows Louis Pujol's thesis defended in 2022.

### 4.6 Railway operations

Following the CIFRE PhD of Rémi Coulaud, we continue our ongoing collaboration with SNCF–Transilien to exploit large datasets of railway operation and passenger flows, obtained by automatic recording devices (for passenger flows, these correspond to sensors at the door level). We model and forecast passenger movement inside train coaches, so as to be able to provide incoming passengers with information on how crowded wagons are. We link this problem with a neural network framework to improve performance. The next step is to take into account the behavior of passengers on platform. This will be examined in a new CIFRE PhD contract to start in 2025.

### 4.7 Ancient materials

Celeste collaborates with CNRS-IPANEMA (Ancient Materials Research Platform). The goal is to propose a new image segmentation method based on a dissimilarity which is particularly well adapted to XRF images. This will allow less exposure to radiation, which is important when dealing with antiques.

### 4.8 Education sciences

Ensuring student success is a central goal of universities, and a high success rate is seen as an indicator of the academic excellence of the institution. The transition from high school to university is seen as a critical time for first-time students, who must adapt not only to a different academic environment but also to greater autonomy. Universities face the challenge of facilitating this transition for a student population that is heterogeneous in terms of academic preparation, cultural and socio-economic background.

We currently collaborate with the EST laboratory (Univ. Paris-Saclay) on this topic. Our works are aimed at identifying the different factors that hinder success, identifying success profiles, and proposing welcoming and support solutions to effectively encourage the success of each student [14, 25].

## 5 Social and environmental responsibility

### 5.1 Footprint of research activities

The carbon emissions of Celeste team members related to their jobs were very low and came essentially from:

- limited levels of transport to and from work, and a small amount for essentially land travel to conferences in France and Europe.
- electronic communication (email, Google searches, Zoom meetings, online seminars, etc.).
- the carbon emissions embedded in their personal computing devices (construction), either laptops or desktops.
- electricity for personal computing devices and for the workplace, plus also water, heating, and maintenance for the latter. Note that only 7.1% (2018) of France's electricity is not sourced from nuclear energy or renewables so team member carbon emissions related to electricity are minimal.

In terms of magnitude, the largest per capita ongoing emissions (excluding flying) are likely simply to be those from buying computers that have a carbon footprint from their construction, in the range of 100 kg Co2-e each. In contrast, typical email use per year is around 10 kg Co2-e per person, and a Zoom call comes to around 10g Co2-e per hour per person, while web browsing uses around 100g Co2-e per hour. Consequently, 2024 was a low carbon year for the Celeste team.

The approximate (rounded for simplicity) kg Co2-e values cited above come from the book, "How Bad are Bananas" by Mike Berners-Lee (2020) which estimates carbon emissions in everyday life.

### 5.2 Impact of research results

In addition to the long-term impact of our theoretical work—which is of course impossible to assess immediately—we are involved in several applied research projects which aim at having a short/mid-term positive impact on society.

First, we collaborate with the EST laboratory (Univ. Paris-Saclay) on questions related to student success in universities and how to maximize it (see Section 4.8).

Second, we collaborate with the Pasteur Institute on neglected tropical diseases; encephalitis in particular, with implications in global health strategies.

Third, the broad use of artificial intelligence/machine learning/statistics nowadays comes with several major ethical issues, one being to avoid making unfair or discriminatory decisions. Our theoretical work on algorithmic fairness has already led to several "fair" algorithms that could be widely used in the short term (one of them is already used for enforcing fair decision-making in student admissions at the University of Genoa).

Fourth, we expect short-term positive impact on society from several direct collaborations with companies such as EDF (forecasting and control of electricity load consumption, in particular, for electric vehicles) and Metafora (early detection of cancers).

## 6 Highlights of the year

### 6.1 Awards

- C. Keribin was promoted to full professor

## 7 New software, platforms, open data

### 7.1 New software

#### 7.1.1 FedCP-QQ

**Name:** Federated Conformal Prediction with Quantile-of-Quantiles

**Keywords:** Prediction set, Conformal prediction, Federated learning, Differential privacy

**Functional Description:** Code of the methods Federated Conformal Prediction with Quantile-of-Quantiles (FedCP-QQ) and its differentially-private version FedCP$^2$-QQ proposed and studied by [41], for building prediction intervals in a one-shot federated learning setting.

**URL:** https://github.com/pierreHmbt/FedCP-QQ

**Contact:** Pierre Humbert

## 8 New results

### 8.1 A tribute to Le Cam: On the van Trees inequality

**Participants:** Gilles Stoltz.

In [11], with Elisabeth Gassiat (Orsay) and in honor of the 100th birth anniversary of Lucien Le Cam (November 18, 1924—April 24, 2000), we developed a version of the van Trees (1968) inequality in the spirit of Hajek–Le Cam, i.e., under minimal assumptions that, in particular, involve no direct point-wise regularity assumptions on densities but rather almost-everywhere differentiability in the model's quadratic mean. Surprisingly, it suffices that the latter differentiability holds along canonical directions—not along all directions. Also, we identified a (slightly stronger) version of the van Trees

inequality as an actual instance of a Cramér-Rao bound, i.e., the van Trees inequality is not just a Bayesian analog of the Cramér–Rao bound, as believed so far. We provided, as an illustration, an elementary proof of the local asymptotic minimax theorem for quadratic loss functions, again assuming differentiability in quadratic mean only along canonical directions.

## 8.2 Marginal and training-conditional guarantees in one-shot federated conformal prediction

**Participants:** Pierre Humbert, Sylvain Arlot.

In collaboration with Batiste Le Bars and Aurélien Bellet (Inria Lille, Magnet project-team), we study in [32] conformal prediction in the one-shot federated learning setting. The main goal is to compute marginally and training-conditionally valid prediction sets, at the server-level, in only one round of communication between the agents and the server. Using the quantile-of-quantiles family of estimators and split conformal prediction, we introduce a collection of computationally-efficient and distribution-free algorithms that satisfy the aforementioned requirements. Our approaches come from theoretical results related to order statistics and the analysis of the Beta-Beta distribution. We also prove upper bounds on the coverage of all proposed algorithms when the nonconformity scores are almost surely distinct. For algorithms with training-conditional guarantees, these bounds are of the same order of magnitude as those of the centralized case. Remarkably, this implies that the one-shot federated learning setting entails no significant loss compared to the centralized case. Our experiments confirm that our algorithms return prediction sets with coverage and length similar to those obtained in a centralized setting.

## 8.3 First-order ANIL provably learns representations despite over-parametrization

**Participants:** Etienne Boursier.

In collaboration with Oguz Yüksel and Nicolas Flammarion (EPFL), we give in [24] convergence guarantees for the well known meta-learning algorithm ANIL in the overparametrized regime. More precisely, we show in the limit of an infinite number of tasks, that first-order ANIL with an overparametrized linear two-layer network architecture successfully learns linear shared representations. The learned solution then yields a good adaptive performance on any new task after a single gradient step. Overall, this illustrates how well model-agnostic methods such as first-order ANIL can learn shared representations.

## 8.4 Estimating the history of a random recursive tree

**Participants:** Simon Briend, Christophe Giraud.

In collaboration with Gabor Lugosi and Déborah Sulen, we study in [9] the problem of estimating the order of arrival of the vertices in a random recursive tree. Specifically, we study two fundamental models: the uniform attachment model and the linear preferential attachment model. We propose an order estimator based on the Jordan centrality measure and define a family of risk measures to quantify the quality of the ordering procedure. Moreover, we establish a minimax lower bound for this problem, and prove that the proposed estimator is nearly optimal. Finally, we numerically demonstrate that the proposed estimator outperforms degree-based and spectral ordering procedures.

### 8.5 Minimax optimal seriation in polynomial time

**Participants:** Christophe Giraud.

In collaboration with Yann Issartel and Nicolas Verzelen, we investigate in [33] the statistical seriation problem, where the statistician seeks to recover a hidden ordering from a noisy observation of a permuted Robinson matrix. In this paper, we tightly characterize the minimax rate for this problem of matrix reordering when the Robinson matrix is bi-Lipschitz, and we also provide a polynomial time algorithm achieving this rate; thereby answering two open questions of [Giraud et al., 2021]. Our analysis further extends to broader classes of similarity matrices.

### 8.6 Computation-information gap in high-dimensional clustering

**Participants:** Bertrand Even, Christophe Giraud.

In collaboration with Nicolas Verzelen, we investigate in [21] the existence of a fundamental computation-information gap for the problem of clustering a mixture of isotropic Gaussian in the high-dimensional regime, where the ambient dimension $d$ is larger than the number $n$ of points. We provide evidence of the existence of such a gap generically in the high-dimensional regime $d \geq n$, by (i) proving a non-asymptotic low-degree polynomials computational barrier for clustering in high-dimension, matching the performance of the best known polynomial time algorithms, and by (ii) establishing that the information barrier for clustering is smaller than the computational barrier, when the number $K$ of clusters is large enough. These results are in contrast with the (moderately) low-dimensional regime $n \geq poly(d,K)$, where there is no computation-information gap for clustering a mixture of isotropic Gaussian. In order to prove our low-degree computational barrier, we develop sophisticated combinatorial arguments to upper-bound the mixed moments of the signal under a Bernoulli Bayesian model.

### 8.7 Active clustering with bandit feedback

**Participants:** Christophe Giraud.

In collaboration with Victor Thuot, Alexandra Carpentier, and Nicolas Verzelen, we investigate in [38] the Active Clustering Problem (ACP). A learner interacts with an $N$-armed stochastic bandit with $d$-dimensional subGaussian feedback. There exists a hidden partition of the arms into $K$ groups, such that arms within the same group, share the same mean vector. The learner's task is to uncover this hidden partition with the smallest budget - i.e., the least number of observation - and with a probability of error smaller than a prescribed constant $\delta$. In this paper, (i) we derive a non-asymptotic lower bound for the budget, and (ii) we introduce the computationally efficient ACB algorithm, whose budget matches the lower bound in most regimes. We improve on the performance of a uniform sampling strategy. Importantly, contrary to the batch setting, we establish that there is no computation-information gap in the active setting.

### 8.8 Policy optimization for adversarial Markov decision processes

**Participants:** Daniil Tiapkin, Evgenii Chzhen, Gilles Stoltz.

In [39], we consider policy optimization in adversarial $H$-episodic Markov decision processes (MDPs) and in the full information setting, Briefly, the learning agent interacts with an environment during $T$

episodes, each of which consists of $H$ stages, and each episode is evaluated with respect to a reward function that will be revealed only at the end of the episode; transition kernels are constant over episodes but unknown. The best regret known so far was of order $S\sqrt{AT}$, as far as the orders of magnitude of $T$ and the sizes $S$ and $A$ of the state and action spaces are concerned. We propose an algorithm, called APO-MVP, that achieves a regret bound of order $\sqrt{SAT}$, matching the minimax lower bound. The proposed algorithm leverages and combines two recent techniques: policy optimization based on online linear optimization strategies (Jonckheere et al., 2023) and a refined martingale analysis of the impact on values of estimating transitions kernels (Zhang et al., 2023).

## 8.9 Incentivized learning in principal-agent bandit games

**Participants:** Etienne Boursier.

In collaboration with A. Scheid (Polytechnique), D. Tiapkiin (LMO), A. Capitaine (Polytechnique), E-M El Mhamdi (Polytechnique), E. Moulines (Polytechnique), M. Jordan (Inria Paris, UC Berkeley), and A. Durmus (Polytechnique), we study in [23] a repeated principal-agent bandit game, where the principal can only interact with her environment through the agent. The principal and the agent have misaligned objectives and the choice of action is only left to the agent. However, the principal can influence the agent's decisions by offering incentives which add up to her rewards. The principal aims to iteratively learn an incentive policy to maximize her own total utility. This framework extends usual bandit problems and is motivated by several practical applications, such as healthcare or ecological taxation, where traditionally used mechanism design theories often overlook the learning aspect of the problem. We present nearly optimal learning algorithms for the principal's regret in both the multi-armed and linear contextual settings.

## 8.10 Learning to mitigate externalities: the Coase Theorem with hindsight rationality

**Participants:** Etienne Boursier.

In collaboration with A. Scheid (Polytechnique), A. Capitaine (Polytechnique), E. Moulines (Polytechnique), M. Jordan (Inria Paris, UC Berkeley), and A. Durmus (Polytechnique), we extend in [22] the result from Section 8.9 to the case where both the agent and principal are learning. In particular, we design a principal strategy that has low regret guarantees against any agent following a learning strategy inside a class of suitable no-regret strategies.

## 8.11 A survey on multi-player bandits

**Participants:** Etienne Boursier.

In collaboration with V. Perchet (ENSAE), we give in [8] a broad overview of recent developments that have appeared over the past decade to do with multiplayer bandits. Multiplayer bandits have recently been extensively studied because of their application to cognitive radio networks. Although considerable progress has been made on theoretical aspects, current algorithms are far from applicable and many obstacles remain between these theoretical results and a possible implementation of multiplayer bandits algorithms in real cognitive radio networks. This survey contextualizes and organizes the rich multiplayer bandits literature.

## 8.12 Unravelling in collaborative learning

**Participants:** Etienne Boursier.

In collaboration with A. Capitaine (Polytechnique), A. Scheid (Polytechnique), E. Moulines (Polytechnique), M. Jordan (Inria Paris, UC Berkeley), E-M El Mhamdi (Polytechnique), and A. Durmus (Polytechnique), we study in [19] the problem of collaborative learning with strategic agents. When such agents wish to train a model together but have sampling distributions of different quality, the coalition may undergo a phenomenon known as unravelling, wherein it shrinks up to the point that it becomes empty or solely comprised of the worst agent. We address this issue by proposing a novel method inspired by probabilistic verification. This approach makes the grand coalition a Nash equilibrium with high probability, despite information asymmetry, thereby hindering unravelling.

## 8.13 Statistical Learning Tools for Electricity Load Forecasting

**Participants:** Jean-Michel Poggi.

The monograph [26], written with Anestis Antoniadis (Institute of Applied Sciences and Intelligent Systems 'Eduardo Caianiello', Naples), Jairo Cugliari (Lab. ERIC EA 3083, Lumière University Lyon 2), Matteo Fasiolo (School of Mathematics, University of Bristol) and Yannig Goude (EDF R& D & LMO), explores a set of statistical and machine learning tools that can be effectively utilized for applied data analysis in the context of electricity load forecasting. Drawing on their substantial research and experience with forecasting electricity demand in industrial settings, the authors guide readers through several modern forecasting methods and tools from both industrial and applied perspectives – generalized additive models (GAMs), probabilistic GAMs, functional time series and wavelets, random forests, aggregation of experts, and mixed effects models. A collection of case studies based on sizable high-resolution datasets, together with relevant R packages, then illustrate the implementation of these techniques. Five real datasets at three different levels of aggregation (nation-wide, region-wide, or individual) from four different countries (UK, France, Ireland, and the USA) are utilized to study five problems: short-term point-wise forecasting, selection of relevant variables for prediction, construction of prediction bands, peak demand prediction, and use of individual consumer data.

This text is intended for practitioners, researchers, and post-graduate students working on electricity load forecasting; it may also be of interest to applied academics or scientists wanting to learn about cutting-edge forecasting tools for application in other areas. Readers are assumed to be familiar with standard statistical concepts such as random variables, probability density functions, and expected values, and to possess some minimal modeling experience.

## 8.14 Diversity-preserving stochastic bandits

**Participants:** Gilles Stoltz.

In [13], with Hédi Hadiji (CentraleSupélec), we revisit the bandit-based framework for diversity-preserving recommendations introduced by Celis, Kapoor, Salehi, and Vishnoi ("Controlling polarization in personalization: an algorithmic framework", at FAccT'2019).

Typically, in stochastic bandits, learning strategies focus quickly on the best-performing arm(s), and play only it (them), except for at most logarithmically many rounds. This phenomenom is called polarization. Celis et al. (2019) proposed a model to avoid polarization: by imposing that the player should play in two stages, first by picking a probability distribution over the arms located in a strict subset of the simplex (the set of all probability distributions)—typically, a polytope in the inner of the simplex. This modeling is interesting and we took it as a starting point.

Our contribution is to get sharper theoretical results (lower bounds and refined upper bounds). Celis et al. (2019) approached the problem in the case of a polytope mainly by a reduction to the setting of linear bandits. We instead design a direct UCB strategy using the specific structure of the setting and show that it enjoys a bounded distribution-dependent regret in the natural cases when the optimal mixed actions put some probability mass on all actions (i.e., when diversity is desirable). The regret lower bounds provided show that otherwise (at least when the model is mean-unbounded), a $\ln T$ regret is suffered. We also discuss an example beyond the special case of polytopes.

### 8.15 Regression under demographic parity constraints via unlabeled post-processing

**Participants:** Evgenii Chzhen, Gayane Taturan.

In collaboration with Gayane Taturyan and Mohamed Hebiri, we study in [20] the problem of performing regression while ensuring demographic parity, even without access to sensitive attributes during inference. We present a general-purpose post-processing algorithm that, using accurate estimates of the regression function and a sensitive attribute predictor, generates predictions that meet the demographic parity constraint. Our method involves discretization and stochastic minimization of a smooth convex function. It is suitable for online post-processing and multi-class classification tasks only involving unlabeled data for the post-processing. Unlike prior methods, our approach is fully theory-driven. We require precise control over the gradient norm of the convex function, and thus, we rely on more advanced techniques than standard stochastic gradient descent. Our algorithm is backed by finite-sample analysis and post-processing bounds, with experimental results validating our theoretical findings.

## 9 Bilateral contracts and grants with industry

**Participants:** Alexandre Janon, Christine Keribin, Jean-Michel Poggi, Gilles Stoltz.

### 9.1 Bilateral contracts with industry

- A. Janon: Contract with INSERM Toulouse (3,3 kE), on variable selection for identification of link between microbiotal dysbiosis and type-2 diabetes.
- C. Keribin: Ongoing Cifre PhD contract with Metafora (30 kE) on machine learning in flow cytometry for early detection of cancers.
- J.M. Poggi: Analysis and modelling of NO2 numerical model biases for data fusion of heterogeneous measurement networks, ATMO NORMANDIE, 20 kE.
- J.M. Poggi, G. Stoltz: Participation in the EDF–Inria Grand défi, with in particular a CIFRE PhD started in December 2023 and a post-doc to start in February 2025.
- G. Stoltz: Ongoing contract with BNP Paribas (3 x 10 kE), on stochastic bandits under budget constraints, with applications to loan management; annually since 2021.

## 10 Partnerships and cooperations

### 10.1 National initiatives

**Participants:** Sylvain Arlot, Kevin Bleakley, Evgenii Chzhen, Christophe Giraud, Gilles Stoltz.

#### 10.1.1 ANR

Sylvain Arlot, Evgenii Chzhen, Christophe Giraud and Gilles Stoltz are part of the PEPR-IA grant CAUSALI-T-AI (CAUSALIty Teams up with Artificial Intelligence), which is led by Marianne Clausel (Univ. de Lorraine).

Sylvain Arlot and Christophe Giraud are part of the ANR Chair-IA grant Biscotte, which is led by Gilles Blanchard (Université Paris Saclay).

Christophe Giraud is part of the ANR ASCAI: Active and batch segmentation, clustering, and seriation: toward unified foundations in AI, with Potsdam University, Munich University, Montpellier INRAE.

#### 10.1.2 Other

Kevin Bleakley worked until September 2024 at 1/3-time (*disponibilité*) with IRT SystemX under the umbrella of Confiance.AI on the subject of anomaly detection in high-dimensional time series data for French industry.

# 11 Dissemination

**Participants:** Sylvain Arlot, Etienne Boursier, Evgenii Chzhen, Bertrand Even, Christophe Giraud, Alexandre Janon, Christine Keribin, Pascal Massart, Jean-Michel Poggi, Marie-Anne Poursat, Gilles Stoltz.

## 11.1 Promoting scientific activities

### 11.1.1 Scientific events: organisation

**General chair, scientific chair**

- C. Keribin is Vice-President of the French Statistical Society (SFdS); member of the board of MALIA, SFdS specialized group in Machine Learning and AI.
- J.-M. Poggi is Past-President of ENBIS (European Network for Business and Industrial Statistics)

**Member of the organizing committees**

- S. Arlot is member of the scientific committee of the Séminaire Palaisien
- E. Chzhen is co-organizer of the DATAIA seminar
- A. Janon is co-organizer the of UQSay seminar
- C. Keribin was co-organizer of the Frugalias workshop (4/10/2024)
- J.-M. Poggi is Member of the ENBIS Award Committee for George Box Medal 2024

### 11.1.2 Scientific events: selection

**Member of the conference program committees**

- C. Giraud was Area Chair for COLT 2024

**Reviewer**

- We performed many reviews for various international conferences.

### 11.1.3 Journal

- We performed many reviews for various international journals.

**Member of the editorial boards**

- S. Arlot: Associate editor for *Annales de l'Institut Henri Poincaré B – Probability and Statistics*
- C. Giraud: Action Editor for JMLR
- C. Giraud: Associate Editor for ESAIM-proc
- C. Keribin: member of the editorial board, *Statistique et Société* (SFdS).
- P. Massart: Associate editor for Panoramas et Synthèses (SMF), Foundations and Trends in Machine Learning, and Confluentes Mathematici
- J.-M. Poggi is Associate Editor for Advances in Data Analysis and Classification
- J.-M. Poggi is Associate Editor for JDSSV J. Data Science, Statistics and Visualization
- J.-M. Poggi was Guest editor for the Springer-Nature Book "Methodological and Applied Statistics and Demography" with the selected papers of this Conference SIS 2024.
- J.-M. Poggi was Guest Editor of the Special Issue on Digital Twins for the Applied Stochastic Models in Business and Industry journal
- G. Stoltz: associate editor for *Mathematics of Operations Research*

**Reviewer - reviewing activities**

- We performed many reviews for various international journals.

#### 11.1.4 Invited talks

- C. Giraud, Mathematical Aspects of Learning Theory - 20 years later, Barcelona (Spain), 9-13 September 2024
- C. Giraud, Colloquium université de Tours
- B. Even, CIRM, Luminy, 16-20 December 2024.
- C. Keribin, Séminaire parisien de statistique, Model Based Co-Clustering: High Dimension and
- J.-M. Poggi, Invited conference, Statistics and Machine Learning in Industry: combining heterogeneous or multi-scale model outputs, CISEM 2024, Mahdia (Tunisia), 17-19 May, 2024
- J.-M. Poggi, Invited conference, Statistics and Machine Learning in Industry: combining heterogeneous or multi-scale model outputs, SIS 2024, Bari (Italy), 17-20 June, 2024 Estimation Challenges, Paris (11/03/2024)
- G. Stoltz, seminar at ESSEC, June 2024

#### 11.1.5 Research administration

- S. Arlot is a member of the council of the Computer Science Graduate School (GS ISN) of University Paris-Saclay.
- S. Arlot is a member of the council of the Computer Science Doctoral School (ED STIC) of University Paris-Saclay.
- C. Giraud is a member of the Scientific Committee of labex IRMIA+, Strasbourg.
- C. Giraud is deputy director of the Mathematics Graduate School of University Paris-Saclay.
- C. Giraud is in charge of the whole Masters program in mathematics for University Paris-Saclay.

- C. Giraud is a member of the local Scientific Committee of Institut Pascal.
- C. Giraud is a member of the council of the Mathematics Doctoral School (EDMH) of Université Paris-Saclay.
- C. Keribin is member of the board of the Computer Science Doctoral School (ED MSTIC) of Paris-Est Sup.
- C. Keribin is Vice-president of the Math - CCUPS (Commission consultative de l'Université Paris–Saclay).
- C. Keribin is member of the council of the mathematics department.
- C. Keribin is in charge of the M2-Math and IA program master of the mathematical school
- P. Massart is director of the Fondation Mathématique Jacques Hadamard.
- M-A. Poursat is in charge of the M1-Mathematics and artificial intelligence program in the master of the mathematical school

#### 11.1.6 Service to the academic community

- Kevin Bleakley: Maintains the English version of the LMO's website dedicated to research activities
- E. Boursier: member of Inria Saclay scientific committee
- E. Chzhen: member of Bibliothèque Jacques Hadamard scientific committee
- C. Giraud: coordinator of computing resources at the Institut Mathématiques d'Orsay (10 engineers)
- C. Giraud: senior member of CCUPS (Commission Consultative Université Paris Saclay)
- C. Giraud is in charge of the Reconvert-AI program.
- C. Keribin is co-president of the scholarship allocation committee MixtAI of the SaclAI school.
- C. Keribin is member of the committee for awarding the Sophie Germain excellence scholarships (FMJH)
- C. Keribin: member of the follow-up committee for PhD student Sara Madad (UTT)
- C. Keribin: member of the follow-up committee for PhD student Anderson Augusma (Laboratoire d'informatique de Grenoble)
- C. Keribin: member of the follow-up committee for PhD student Augustin Pion (Laboratoire des Signaux et Systèmes, CentraleSupelec)

### 11.2 Teaching - Supervision - Juries

#### 11.2.1 Teaching

Most of the team members (especially Professors, Associate Professors and Ph.D. students) teach several courses at University Paris-Saclay, as part of their teaching duty. We mention below some of the classes in which we teach.

- Masters: S. Arlot, Statistical learning and resampling, 30h, M2, Université Paris-Sud
- Masters: S. Arlot, Preparation for French mathematics agrégation (statistics), 25h, M2, Université Paris-Sud
- Masters: E. Boursier, Sequential Learning, 24h, M2 Université Paris-Saclay
- Licence/Masters: E. Chzhen, PCC Polytechnique

- Masters: E. Chzhen, Statistical Theory of Algorithmic Fairness, 20h, M2 Université Paris-Saclay
- Masters: C. Giraud, High-Dimensional Probability and Statistics, 45h, M2, Université Paris-Saclay
- Masters: C. Giraud, Mathematics for AI, 75h, M1, Université Paris-Saclay
- Masters: C. Keribin, unsupervised and supervised learning, M1, 42h, Université Paris-Saclay
- Masters: C. Keribin, Cours accéléré en statistiques, M2, 21h, Université Paris-Saclay
- Masters: C. Keribin, Modélisation statistique, M1, 40h, Université Paris-Saclay
- Masters: C. Keribin, Advanced Unsupervised Learning, M2, 24h, Université Paris-Saclay
- Masters: C. Keribin, Internship supervision for M1-Applied Mathematics and M2-DataScience, Université Paris-Saclay
- Masters: M-A Poursat, applied statistics, 21h, M1 Artificial Intelligence, Université paris-Saclay
- Masters: M-A Poursat, statistical learning, 42h, M2 Bioinformatics, Université paris-Saclay
- Masters: M-A Poursat, méthodes de classification, 24h, M1, Université Paris Saclay
- Licence: M-A Poursat, inférence statistique, 72h, L3, Université Paris Saclay
- Masters: G. Stoltz, Introduction to data science with Python, 18h, M1 HEC Paris

#### 11.2.2 Supervision

- PhD defended on January 12, 2024: Karl Hajjar, A dynamical analysis of infinitely wide neural networks, started Oct. 2020, C. Giraud and L. Chizat.
- PhD defended on June 11, 2024: Simon Briend, Inference of the past of random structures and other random problems, started Sept. 2021, co-advised by C. Giraud and G. Lugosi
- PhD in progress: Samy Clementz, Data-driven Early Stopping Rules for saving computation resources in AI, started Sept. 2021, co-advised by S. Arlot and A. Celisse
- PhD in progress: Gayane Taturyan, Fairness and Robustness in Machine Learning, started Nov. 2021, co-advised by E. Chzhen, J.-M. Loubes (Univ. Toulouse Paul Sabatier) and M. Hebiri (Univ. Gustave Eiffel)
- PhD in progress: Leonardo Martins-Bianco, Disentangling the relationships between different community detection algorithms, started October 2022, co-advised by C. Keribin and Z. Naulet (Univ. Paris-Saclay)
- PhD in progress: Chiara Mignacco, Aggregation (orchestration) of reinforcement learning policies, started October 2022, co-advised by G. Stoltz and Matthieu Jonckheere (LAAS Toulouse)
- PhD in progress: Pierre-André Mikem, Multiple instance learning for the detection of tumor cells, started March 2023, co-advised by C. Keribin and P. Massart (Univ. Paris-Saclay). Cifre contract with Metafora.
- PhD in Progress: Aymeric Capitaine, Incitivizing Federated and Decentralized Learning, started September 2023, co-advised by E. Boursier, M. Jordan (Inria Paris) and A. Durmus (Polytechnique)
- PhD in Progress: Antoine Scheid, Multi-agent bandits and Markovian games, started September 2023, co-advised by E. Boursier, M. Jordan (Inria Paris) and A. Durmus (Polytechnique)
- PhD in Progress: Daniil Tipakin, Topics about sample complexity in reinforcement learning, started October 2023, co-advised by G. Stoltz and E. Moulines (Polytechnique)

- PhD in Progress: Guillaume Principato, Hierarchical conformal prediction for smart electric vehicle charging, started December 2023, co-advised by J.M. Poggi and G. Stoltz, as well as Y. Amara-Ouali, Y. Goude, B. Hamrouche (EDF)
- PhD in progress: Bertrand Even, Compromis Statistique-Computationnel et équité en apprentissage non-supervisé, started September 2024, co-advised by C. Giraud and N. Verzelen (Inrae)
- PhD in progress: Dhia-Elhaq Ouerfelli, Change-point detection and explainability of high-dimensional time series, started October 2024, co-advised by S. Arlot, K. Bleakley, and P. Pamphile
- PhD in progress: Guillermo Martin, Tirer partie de l'IA explicable pour améliorer les prévisions des chaînes d'approvisionnement, started October 2024, co-advised by C. Giraud and O. Klopp (ESSEC Business School)
- PhD in progress: Victor Turmel, Repeated Games and Sequential Learning: Towards Fair and Efficient Algorithms, started October 2024, co-advised by G. Stoltz and E. Boursier

#### 11.2.3 Juries

We participated in many PhD committees (too many to keep an exact record), at University Paris-Saclay as well as at other universities, and we refereed several of these PhDs.

### 11.3 Popularization

#### 11.3.1 Education

Christophe Giraud produces educational videos on his YouTube channel "High-dimensional probability and statistics".

Gilles Stoltz holds MATh.en.JEANS workshops in secondary schools in Laval (collège Fernand Puech, Lycée Douanier Rousseau).

Patrick Pamphile is a scientific tutor for scientific discovery projects in middle schools, proposed by the F93 association. In 2024, he did 20 hours of lectures on population dynamics at Collège Théodore Monod, 93220 Gagny.

#### 11.3.2 Interventions

- Christine Keribin organises and chairs a session on professions in mathematics, statistics and AI during the Forum Emploi Maths.

## 12 Scientific production

### 12.1 Major publications

[1] A. Antoniadis, J. Cugliari, M. Fasiolo, Y. Goude and J.-M. Poggi. *Statistical Learning Tools for Electricity Load Forecasting*. Statistics for Industry, Technology, and Engineering. Springer International Publishing, 2024. DOI: 10.1007/978-3-031-60339-6. URL: https://hal.science/hal-04673275.

[2] E. Boursier and V. Perchet. 'A survey on multi-player bandits'. In: *Journal of Machine Learning Research* (Jan. 2024). URL: https://inria.hal.science/hal-03941302.

[3] E. Chzhen, M. Hebiri and G. Taturyan. 'Regression under demographic parity constraints via unlabeled post-processing'. In: *PMLR*. NeurIPS 2024. Vancouver, Canada, 9th Dec. 2024. URL: https://hal.science/hal-04654182.

[4] B. Even, C. Giraud and N. Verzelen. 'Computation-information gap in high-dimensional clustering'. In: Proceedings of Thirty Seventh Conference on Learning Theory. Vol. 247. Edmonton (Canada), Canada, 30th June 2024, pp. 1646–1712. URL: https://hal.science/hal-04483306.

[5] É. Gassiat and G. Stoltz. 'The van Trees inequality in the spirit of Hajek and Le Cam'. In: *Statistical Science* (2024). URL: https://hal.science/hal-04452222.

[6] C. Giraud, Y. Issartel, L. Lehericy and M. Lerasle. 'Pair-Matching: Link Prediction with Adaptive Queries'. In: *Mathematical Statistics and Learning* (5th Mar. 2024). URL: https://hal.science/hal-04578273.

[7] O. Yuksel, E. Boursier and N. Flammarion. 'First-order ANIL provably learns representations despite overparametrization'. In: ICLR 2024 - The Twelfth International Conference on Learning Representations. Vienne, Austria, 7th May 2024. URL: https://inria.hal.science/hal-04105211.

## 12.2 Publications of the year

### International journals

[8] E. Boursier and V. Perchet. 'A survey on multi-player bandits'. In: *Journal of Machine Learning Research* (Jan. 2024). URL: https://inria.hal.science/hal-03941302 (cit. on p. 10).

[9] S. Briend, C. Giraud, G. Lugosi and D. Sulem. 'Estimating the history of a random recursive tree'. In: *Bernoulli* (2025). URL: https://hal.science/hal-04884909. In press (cit. on p. 8).

[10] O. Coudray, P. Bristiel, M. Dinis, C. Keribin and P. Pamphile. 'Construction of fatigue criteria through Positive Unlabeled Learning'. In: *Fatigue and Fracture of Engineering Materials and Structures* 48.1 (17th Oct. 2024), pp. 101–117. DOI: 10.1111/ffe.14452. URL: https://inria.hal.science/hal-04324629 (cit. on p. 5).

[11] É. Gassiat and G. Stoltz. 'The van Trees inequality in the spirit of Hajek and Le Cam'. In: *Statistical Science* (2024). URL: https://hal.science/hal-04452222 (cit. on p. 7).

[12] C. Giraud, Y. Issartel, L. Lehericy and M. Lerasle. 'Pair-Matching: Link Prediction with Adaptive Queries'. In: *Mathematical Statistics and Learning* (5th Mar. 2024). URL: https://hal.science/hal-04578273.

[13] H. Hadiji, S. Gerchinovitz, J.-M. Loubes and G. Stoltz. 'Diversity-Preserving K-Armed Bandits, Revisited'. In: *Transactions on Machine Learning Research Journal* July (July 2024). URL: https://hal.science/hal-02957485 (cit. on p. 11).

[14] P. Pamphile and I. Bournaud. 'Analyses Statistiques Exploratoires de Données en Éducation : Principes, Concepts et Implémentation.Le cas de l'adaptation des primo-entrant.es en IUT après la réforme du baccalauréat et du BUT'. In: *Mesure et Evaluation en Education* 47.3 (2024). URL: https://inria.hal.science/hal-04375594 (cit. on p. 6).

[15] A. Sportisse, M. Marbac, F. Laporte, G. Celeux, C. Boyer, J. Josse and C. Biernacki. 'Model-based Clustering with Missing Not At Random Data'. In: *Statistics and Computing* (18th June 2024). DOI: 10.1007/s11222-024-10444-2. URL: https://hal.science/hal-03494674.

### Invited conferences

[16] C. Biernacki, J. Jacques and C. Keribin. 'MODEL BASED CO-CLUSTERING: HIGH DIMENSION & ESTIMATION CHALLENGES'. In: RMR 2024 : Modèles statistiques pour des données dépendantes et applications. Rouen, France, 19th June 2024. URL: https://inria.hal.science/hal-04867840.

[17] C. Biernacki, J. Jacques and C. Keribin. 'MODEL BASED CO-CLUSTERING: HIGH DIMENSION & ESTIMATION CHALLENGES'. In: CFE-CMStatistics 2024 - The 18th International Joint Conference on Computational and Financial Econometrics (CFE) and Computational and Methodological Statistics (CMStatistics). Londres, United Kingdom, 14th Dec. 2024. URL: https://inria.hal.science/hal-04867739.

**International peer-reviewed conferences**

[18] Z. Benomar, E. Chzhen, N. Schreuder and V. Perchet. 'Addressing bias in online selection with limited budget of comparisons'. In: *PMLR*. NeurIPS 2024. Vancouver (BC), Canada, 9th Dec. 2024. URL: https://hal.science/hal-04275550.

[19] A. Capitaine, E. Boursier, A. Scheid, E. Moulines, M. I. Jordan, E.-M. El-Mhamdi and A. Durmus. 'Unravelling in Collaborative Learning'. In: NeurIPS 2024 - The Thirty-Eighth Annual Conference on Neural Information Processing Systems. Vancouver, Canada, 10th Dec. 2024. URL: https://hal.science/hal-04847788 (cit. on p. 11).

[20] E. Chzhen, M. Hebiri and G. Taturyan. 'Regression under demographic parity constraints via unlabeled post-processing'. In: *PMLR*. NeurIPS 2024. Vancouver, Canada, 9th Dec. 2024. URL: https://hal.science/hal-04654182 (cit. on p. 12).

[21] B. Even, C. Giraud and N. Verzelen. 'Computation-information gap in high-dimensional clustering'. In: Proceedings of Thirty Seventh Conference on Learning Theory. Vol. 247. Edmonton (Canada), Canada, 30th June 2024, pp. 1646–1712. URL: https://hal.science/hal-04483306 (cit. on p. 9).

[22] A. Scheid, A. Capitaine, E. Boursier, E. Moulines, M. I. Jordan and A. Durmus. 'Learning to Mitigate Externalities: the Coase Theorem with Hindsight Rationality'. In: NeurIPS 2024 - The Thirty-Eighth Annual Conference on Neural Information Processing Systems. Vancouver, Canada, 10th Dec. 2024. URL: https://hal.science/hal-04847764 (cit. on p. 10).

[23] A. Scheid, D. Tiapkin, E. Boursier, A. Capitaine, E. Mahdi, É. Moulines, M. I. Jordan and A. Durmus. 'Incentivized Learning in Principal-Agent Bandit Games'. In: ICML 2024 - The Forty-First International Conference on Machine Learning. Vienne, Austria, 21st July 2024. URL: https://hal.science/hal-04479761 (cit. on p. 10).

[24] O. Yuksel, E. Boursier and N. Flammarion. 'First-order ANIL provably learns representations despite overparametrization'. In: ICLR 2024 - The Twelfth International Conference on Learning Representations. Vienne, Austria, 7th May 2024. URL: https://inria.hal.science/hal-04105211 (cit. on p. 8).

**Conferences without proceedings**

[25] P. Pamphile, I. Bournaud and C. Clavel. 'Identifier et comprendre les difficultés d'adaptation des primo entrantes à l'université : utilisation d'une méthode mixte quantitative-qualitative avec des méthodes statistiques d'apprentissage automatique'. In: Diversité, Réussite[s] dans l'Enseignement Supérieur (2024). Nantes (France), France, 3rd Apr. 2024. URL: https://hal.science/hal-04489836 (cit. on p. 6).

**Scientific books**

[26] A. Antoniadis, J. Cugliari, M. Fasiolo, Y. Goude and J.-M. Poggi. *Statistical Learning Tools for Electricity Load Forecasting*. Statistics for Industry, Technology, and Engineering. Springer International Publishing, 2024. DOI: 10.1007/978-3-031-60339-6. URL: https://hal.science/hal-04673275 (cit. on p. 11).

**Doctoral dissertations and habilitation theses**

[27] S. Briend. 'Inference of the past of random structures and other random problems'. Université Paris-Saclay, 11th June 2024. URL: https://theses.hal.science/tel-04653882.

[28] K. Hajjar. 'A dynamical analysis of infinitely wide neural networks'. Université Paris-Saclay, 12th Jan. 2024. URL: https://theses.hal.science/tel-04548479.

**Reports & preprints**

[29] A. Barbier-Chebbah, C. L. Vestergaard, J.-B. Masson and E. Boursier. *Approximate information maximization for bandit games*. 29th Nov. 2024. DOI: 10.48550/arXiv.2310.12563. URL: https://hal.science/hal-04246907.

[30] K. Bleakley. *Extreme change-point detection*. 5th Apr. 2024. URL: https://inria.hal.science/hal-04523912.

[31] S. Gaucher, G. Blanchard and F. Chazal. *Supervised Contamination Detection, with Flow Cytometry Application*. 5th Apr. 2024. URL: https://hal.science/hal-04535142.

[32] P. Humbert, B. Le Bars, A. Bellet and S. Arlot. *Marginal and training-conditional guarantees in one-shot federated conformal prediction*. 18th May 2024. URL: https://hal.science/hal-04579882 (cit. on p. 8).

[33] Y. Issartel, C. Giraud and N. Verzelen. *Minimax optimal seriation in polynomial time*. 14th May 2024. URL: https://hal.science/hal-04575332 (cit. on p. 9).

[34] P. Lacroix, M. Gallopin and M.-L. Martin. *An overview of variable selection procedures using regularization paths in high-dimensional Gaussian linear regression*. 14th Feb. 2024. URL: https://hal.science/hal-03366851.

[35] P. Lacroix and M.-L. Martin. *Supplementary file for the article "Trade-off between predictive performance and FDR control for high-dimensional Gaussian model selection"*. 25th June 2024. URL: https://hal.science/hal-04625023.

[36] P. Lacroix and M.-L. Martin. *Trade-off between predictive performance and FDR control for high-dimensional Gaussian model selection*. 11th Apr. 2024. URL: https://hal.science/hal-03978309.

[37] E. M. Saad, G. Blanchard and S. Arlot. *Online Orthogonal Matching Pursuit*. 7th Oct. 2024. URL: https://hal.science/hal-03141061.

[38] V. Thuot, A. Carpentier, C. Giraud and N. Verzelen. *Active clustering with bandit feedback*. June 2024. URL: https://hal.science/hal-04610780 (cit. on p. 9).

[39] D. Tiapkin, E. Chzhen and G. Stoltz. *Narrowing the Gap between Adversarial and Stochastic MDPs via Policy Optimization*. 5th July 2024. URL: https://hal.science/hal-04636422 (cit. on p. 9).

**Other scientific publications**

[40] C. Keribin, C. Biernacki and J. Jacques. *Model Based Co-Clustering: High Dimension and Estimation Challenges*. 11th Mar. 2024. URL: https://inria.hal.science/hal-04862826.

## 12.3 Cited publications

[41] P. Humbert, B. Le Bars, A. Bellet and S. Arlot. 'One-shot federated conformal prediction'. In: *International Conference on Machine Learning*. PMLR. 2023, pp. 14153–14177 (cit. on p. 7).