RESEARCH CENTRE

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

IN PARTNERSHIP WITH:

CNRS, Université Paris-Saclay

2023
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

Ínria

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, from Apr 2023]

- Gilles Celeux [INRIA, Emeritus]

- Evgenii Chzhen [CNRS, Researcher]

- Gilles Stoltz [CNRS, Senior Researcher, HDR]

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, Associate Professor, HDR]

- Pascal Massart [UNIV PARIS SACLAY, Professor]

- Patrick Pamphile [UNIV PARIS SACLAY, Associate Professor]

- Marie-Anne Poursat [UNIV PARIS SACLAY, Associate Professor]

Post-Doctoral Fellow

- Pierre Humbert [INRIA, Post-Doctoral Fellow]

PhD Students

- Emilien Baroux [GROUPE PSA, until Jun 2023]

- Antoine Barrier [ENS DE LYON, until Aug 2023]

- Samy Clementz [SORBONNE UNIVERSITE]

- Karl Hajjar [UNIV PARIS SACLAY, until Sep 2023]

- Leonardo Martins Bianco [UNIV PARIS-SACLAY]

- Chiara Mignacco [UNIV PARIS SACLAY]

- Pierre-Andre Mikem [UNIV PARIS SACLAY, from Mar 2023]

- Guillaume Principato [EDF, from Dec 2023]

- Gayane Taturyan [IRT SYSTEM X]

- Daniil Tiapkin [Ecole Polytechnique, from Oct 2023]

Interns and Apprentices

- Bertrand Even [ENS, Intern, from Sep 2023]

- Raphael Walker [INRIA, Intern, from Apr 2023 until Sep 2023]

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, 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 (stochastic bandits and expert aggregation), 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 Neglected tropical diseases

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

4.2 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. We currently work on an approach involving back and forth disaggregation (of the total consumption into the consumption of well-chosen groups/regions) and aggregation of local estimates.

4.3 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.

CELESTE has established a collaboration with the manufacturer Stellantis, which has led to the completion of two CIFRE theses. The first was defended in 2022, followed by Emilien Baroux's thesis, which was defended in 2023. In the case of personal vehicles, various loads need to be taken into account due to different road types (e.g. freeway, city) and driving styles (aggressive, sporty, etc.). In Emilien Baroux's thesis we used a multidimensional characterization of damage caused by multiple-input external wheel loads during vehicle use, including load combinations between the left and right wheels of the front and rear axles. Field measurements were used to calculate pseudo-damage for each load and road type, creating multivariate data with a hierarchical structure. Unsupervised analyses were used to explore correlations between pseudo-damage and to identify driving profiles, providing a multidimensional assessment of severity while avoiding over-learning. A multidimensional Gaussian mixture model was then fitted to damage-equivalent constraints. The resulting probabilistic model could then be used to extrapolate damage calculations and simulate driving styles, providing design teams with a stress analysis tool for accurate and realistic fatigue design of chassis components in future vehicles.

4.4 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, in the context of Pierre-André Mikem's Cifre PhD, which follows Louis Pujol's thesis defended in 2022.

4.5 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 support the success of each student.

5   Social and environmental responsibility

5.1   Footprint of research activities

Still influenced by the aftermath of the Covid-19 pandemic, 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 CO_2-e each. In contrast, typical email use per year is around 10 kg CO_2-e per person, and a Zoom call comes to around 10g CO_2-e per hour per person, while web browsing uses around 100g CO_2-e per hour. Consequently, 2023 was a very low carbon year for the Celeste team. To put this in the context of work travel by flying, one return Paris-Nice flight corresponds to 160 kg CO_2-e emissions, which likely dwarfs the total emissions of any one Celeste team member's work-related emissions in 2023.

The approximate (rounded for simplicity) kg CO_2-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 Pasteur Institute on neglected tropical diseases; encephalitis in particular, with implications in global health strategies.

Second, we collaborate with the EST laboratory (Univ. Paris-Saclay) on questions related to student success in universities and how to maximize it (see Section 4.5).

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), Stellantis (automobile reliability, with two Cifre contracts) and Metafora (early detection of cancers).

6   New software, platforms, open data

6.1   New software

6.1.1   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 [19], for building prediction intervals in a one-shot federated learning setting.

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

Contact: Pierre Humbert

7 New results

7.1 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 introduce in [19] a conformal prediction method to construct prediction sets in a one-shot federated learning setting. More specifically, we defined a quantile-of-quantiles estimator and proved that for any distribution, it is possible to output prediction sets with desired coverage in only one round of communication. To mitigate privacy issues, we also described a locally differentially private version of our estimator. Finally, over a wide range of experiments, we showed that our method returns prediction sets with coverage and length very similar to those obtained in a centralized setting. Overall, these results demonstrate that our method is particularly well-suited to perform conformal predictions in a one-shot federated learning setting.

7.2 Parameter-free projected gradient descent

Participants: Evgenii Chzhen, Christophe Giraud, Gilles Stoltz.

In [28], we consider the problem of minimizing a convex function over a closed convex set, with Projected Gradient Descent (PGD). We propose a fully parameter-free version of AdaGrad, which is adaptive to the distance between the initialization and the optimum, and to the sum of the square norm of the subgradients. Our algorithm is able to handle projection steps, does not involve restarts, reweighing along the trajectory or additional gradient evaluations compared to the classical PGD. It also fulfills optimal rates of convergence for cumulative regret up to logarithmic factors. We provide an extension of our approach to stochastic optimization.

7.3 Gradient-free optimization of highly smooth functions: improved analysis and a new algorithm

Participants: Evgenii Chzhen.

In collaboration with Arya Akhavan (CMAP, IP Paris), Massimiliano Pontil (Italian Institute of Technology) and Alexandre Tsybakov (CREST, IP Paris), we study in [24] minimization problems with zero-order noisy oracle information under the assumption that the objective function is highly smooth and possibly satisfies additional properties. We consider two kinds of zero-order projected gradient descent algorithms, which differ in the form of the gradient estimator. The first algorithm uses a gradient estimator based on randomization over the ℓ_2 sphere due to Bach and Perchet (2016). We present an improved analysis of this algorithm on the class of highly smooth and strongly convex functions studied in the prior work, and we derive rates of convergence for two more general classes of non-convex functions.

7.4 SignSVRG: fixing signSGD via variance reduction

Participants: Evgenii Chzhen.

In collaboration with Sholom Schechtman (IP Paris), we consider in [29] the problem of unconstrained minimization of finite sums of functions. We propose a simple, yet, practical way to incorporate variance reduction techniques into SignSGD, guaranteeing convergence that is similar to the full sign gradient descent. The core idea is first instantiated on the problem of minimizing sums of convex and Lipschitz functions and is then extended to the smooth case via variance reduction. Our analysis is elementary and much simpler than the typical proof for variance reduction methods.

7.5 Small Total-Cost Constraints in Contextual Bandits with Knapsacks, with Application to Fairness

Participants: Evgenii Chzhen, Christophe Giraud, Gilles Stoltz.

In collaboration with Zhen Li (BNP Paribas), we consider in [16] contextual bandit problems with knapsacks [CBwK], a problem where at each round, a scalar reward is obtained and vector-valued costs are suffered. The learner aims to maximize the cumulative rewards while ensuring that the cumulative costs are lower than some predetermined cost constraints. We assume that contexts come from a continuous set, that costs can be signed, and that the expected reward and cost functions, while unknown, may be uniformly estimated – a typical assumption in the literature. In this setting, total cost constraints had so far to be at least of order $T^{3/4}$, where T is the number of rounds, and were even typically assumed to depend linearly on T. We are however motivated to use CBwK to impose a fairness constraint of equalized average costs between groups: the budget associated with the corresponding cost constraints should be as close as possible to the natural deviations, of order $\sqrt{T}$.

7.6 Symphony of experts: orchestration with adversarial insights in reinforcement learning

Participants: Chiara Mignacco, Gilles Stoltz.

In collaboration with Matthieu Jonckheere (LAAS Toulouse), we consider in [31] a setting of structured reinforcement learning and work on the concept of orchestration, where a (small) set of expert policies guides decision-making. The modeling thereof, with expert policies considered as super-actions, constitutes our first contribution. We then establish value-functions regret bounds for orchestration in the tabular setting by transferring regret-bound results from adversarial settings. We generalize and extend the analysis of natural policy gradient in Agarwal et al. [2021, Section 5.3] to arbitrary adversarial aggregation strategies. We also extend it to the case of estimated advantage functions, providing insights into sample complexity both in expectation and high probability. A key point of our approach lies in its arguably more transparent proofs compared to existing methods.

7.7 Approximate information maximization for bandit games

Participants: Etienne Boursier.

In collaboration with Alex Barbier-Chebbah, Christian Vestergaard and Jean-Baptiste Masson (Institut Pasteur & Inria Paris, EPIMETHEE project-team), we proposed in [25] a new bandits algorithm based on information maximisation principles. More precisely, we propose a new class of bandit algorithms that maximize an approximation to the information of a key variable within the system. To this end, we develop an approximated analytical physics-based representation of an entropy to forecast the information gain of each action and greedily choose the one with the largest information gain.

7.8 Constant or logarithmic regret in asynchronous multiplayer bandits

Participants: Etienne Boursier.

Multiplayer bandits have recently been extensively studied because of their application to cognitive radio networks. While the literature mostly considers synchronous players, radio networks (e.g. for IoT) tend to have asynchronous devices. This motivates the harder, asynchronous multiplayer bandits problem, which was first tackled with an explore-then-commit (ETC) algorithm (see Dakdouk, 2022), with a regret upper-bound in $O(T^{2/3})$. Before even considering decentralization, understanding the centralized case was still a challenge as it was unknown whether getting a regret smaller than $\Omega(T^{2/3})$ was possible.

In collaboration with Hugo Richard (Criteo and Inria Saclay, Mind project-team) and Vianney Perchet (Criteo and Ensae), in [34] we answer positively this question, as a natural extension of UCB exhibits a $O(\sqrt{T\log(T)})$ minimax regret. More importantly, we introduce Cautious Greedy, a centralized algorithm that yields constant instance-dependent regret if the optimal policy assigns at least one player on each arm (a situation that is proved to occur when arm means are close enough). Otherwise, its regret increases as the sum of $\log(T)$ over some sub-optimality gaps. We provide lower bounds showing that Cautious Greedy is optimal in the data-dependent terms. Therefore, we set up a strong baseline for asynchronous multiplayer bandits and suggest that learning the optimal policy in this problem might be easier than thought, at least with centralization.

7.9 Model-Based Co-Clustering: High Dimension and Estimation Challenges

Participants: Christine Keribin.

Model-based co-clustering can be seen as a particularly important extension of model-based clustering. It allows for a significant reduction of both the number of rows (individuals) and columns (variables) of a data set in a parsimonious manner, and also allows interpretability of the resulting reduced data set since the meaning of the initial individuals and features is preserved. Moreover, it benefits from the rich statistical theory for both estimation and model selection. Many works have produced new advances on this topic in recent years.

In collaboration with Christophe Biernacki (Inria Lille, Modal project-team) and Julien Jacques (Univ. Lyon), we offer in [6] a general update of the related literature. In addition, we advocate two main messages, supported by specific research material: (1) co-clustering requires further research to fix some well-identified estimation issues, and (2) co-clustering is one of the most promising approaches for clustering in the (very) high-dimensional setting, which corresponds to the global trend in modern data sets.

7.10 Construction of fatigue criteria through Positive Unlabeled Learning

Participants: Olivier Coudray, Christine Keribin, Patrick Pamphile.

Vehicles reliability is a major issue for automotive manufacturers. In particular, mechanical fatigue is an important preoccupation of the design office. In order to accelerate the development of new mechanical parts, car manufacturers want to rely more on numerical simulation and drastically reduce the number of validation tests on prototypes. To do this, they need efficient fatigue criteria, able to correctly identify critical zones on a numerical model. However, the current fatigue criteria used to post process numerical results fail to correlate well on fatigue test rig.

In [30], in collaboration with Philippe Bristiel and Miguel Dinis (Stellantis), we first propose a probabilistic Dang Van criterion that accounts for the dispersion of fatigue results in a multiaxial setting. We then introduce a fatigue database built upon numerical results and fatigue test reports on automotive chassis components. A novel approach, based on Positive-Unlabeled learning (PU learning), is developed to leverage this source of data and improve the predictivity of the fatigue criterion. The methodology is applied to the fatigue database to illustrate the interest of the approach.

7.11 Education sciences

Participants: Patrick Pamphile.

Quantitative surveys are effective at identifying general trends, but can lack depth. Qualitative surveys, on the other hand, generate rich textual data that provides a deep, individualized understanding of the situation under study. However, processing this data is more complex.

In our recent works [7, 21, 33, 37] in collaboration with the EST laboratory (Univ. Paris-Saclay), we propose a mixed method that uses both unsupervised methods and neural network models to process both quantitative and textual data, in order to take advantage of the articulation between quantitative and qualitative surveys. See also Section 4.5 about our work applied to education sciences.

7.12 Addressing bias in online selection with limited budget of comparisons

Participants: Evgenii Chzhen.

Consider a hiring process with candidates coming from different universities. It is easy to order candidates who have the same background, yet it can be challenging to compare them otherwise. The latter case requires additional costly assessments and can result in sub-optimal hiring decisions. Given an assigned budget, what would be an optimal strategy to select the most qualified candidate?

In [27], in collaboration with Ziyad Benomar (Inria Saclay, Fairplay project-team), Nicolas Schreuder (LIGM) and Vianney Perchet (Criteo and Ensae), we model the above problem by introducing a new variant of the secretary problem in which sequentially observed candidates are split into two distinct groups. For each new candidate, the decision maker observes its rank among already seen candidates from the same group and can access its rank among all observed candidates at some fixed cost. To tackle this new problem, we introduce and study the family of Dynamic Double Threshold (DDT) algorithms. We show that, with well-chosen parameters, their success probability converges rapidly to $1/e$ as the budget grows, recovering the optimal success probability from the usual secretary problem. Finally, focusing on the class of memory-less algorithms, we propose an optimal algorithm in the non-asymptotic regime and show that it belongs to the DDT family when the number of candidates is large.

8   Bilateral contracts and grants with industry

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

8.1 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.

- P. Pamphile: Collaboration contract with Stellantis (with A. Constantinescu, LMS, IP Paris). 95 kE.

- 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.

- 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.

9 Partnerships and cooperations

9.1 National initiatives

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

9.1.1 ANR

Sylvain Arlot is part of the ANR grant FAST-BIG (Efficient Statistical Testing for high-dimensional Models: application to Brain Imaging and Genetics), which is lead by Bertrand Thirion (Inria Saclay, Parietal).

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.

9.1.2 Other

Kevin Bleakley works 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.

10 Dissemination

Participants: Sylvain Arlot, Kevin Bleakley, Etienne Boursier, Gilles Celeux, Evgenii Chzhen, Christophe Giraud, Alexandre Janon, Christine Keribin, Pascal Massart, Patrick Pamphile, Jean-Michel Poggi, Marie-Anne Poursat, Gilles Stoltz.

10.1   Promoting scientific activities

10.1.1   Scientific events: organisation

General chair, scientific chair

- C. Keribin is member of the board and treasurer of the French Statistical Society (SFdS); member of the board of MALIA, SFdS specialized group in Machine Learning and AI.

- J.-M. Poggi is 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 DATAIA institutional seminar.

- C. Giraud is co-organizer with Estelle Kuhn of StatMathAppli Frejus, 2023.

- A. Janon is co-organizer of UQSay seminar.

- C. Keribin and G. Stoltz organized a 3-day event for the 62nd anniversary of Elisabeth Gassiat in Orsay, June 1–3, 2023, with the support of Inria (among others); about 150 participants, see here.

10.1.2   Scientific events: selection

Member of the conference program committees

- A. Janon is member of the program commitee of the Mascot-Num 2023 conference in Le Croisic.

Reviewer

- We performed many reviews for various international conferences.

10.1.3   Journal

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: Associate Editor Advances in Data Analysis and Classification

- J.-M. Poggi: Associate Editor JDSSV J. Data Science, Statistics and Visualization

- J.-M. Poggi: Associate Editor du Journal of Statistical Software

- G. Stoltz: associate editor for *Mathematics of Operations Research*

Reviewer - reviewing activities

- We performed many reviews for various international journals.

10.1.4   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 in charge of the whole Masters program in mathematics for University Paris-Saclay.

- C. Giraud is member of the board of the Mathematics Graduate School of 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 Mathematic 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-DataScience and M2-Math and IA programs in the 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

10.1.5   Service to the academic community

- K. 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 ressources at the Institut Mathématiques d'Orsay (10 ingeneers)

- 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 commitee for awarding the Sophie Germain excellence scholarships (FMJH)

- C. Keribin: member of the follow-up committee for PhD student Tom Guedon (Inrae)

- 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)

10.2   Teaching - Supervision - Juries

10.2.1   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

- Masters: E. Chzhen, Fairness and Privacy in Machine Learning, 18h, M2 ENSAE

- 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, Accelerated course in statistics, M2, 21h, Université Paris-Saclay

- Masters: C. Keribin, Statistical modelization, 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, Classification methods, 24h, M1, Université Paris Saclay

- Licence: M-A Poursat, Statistical inference, 72h, L3, Université Paris Saclay

- Masters: G. Stoltz, Introduction to data science with Python, 18h, M1 HEC Paris

10.2.2   Supervision

- Ph.D. defended on June 2023: Emilien Baroux, Reliability dimensioning under complex loads: from specification to validation, started July. 2020, co-advised by A. Constantinescu (LMS, IP Paris) and P. Pamphile, CIFRE with STELLANTIS

- Ph.D. defended on July 2023: Antoine Barrier, Contributions to a theory of pure exploration in sequential statistics, started September 2020, co-advised by G. Stoltz and A. Garivier (ENS Lyon)

- PhD in progress: Karl Hajjar, A dynamical analysis of infinitely wide neural networks, started Oct. 2020, C. Giraud and L. Chizat (EPFL).

- 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 (Univ. Paris 1 Panthéon-Sorbonne)

- 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 E.M. El Mahdi (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)

10.2.3   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.

10.3   Popularization

10.3.1   Articles and contents

Christine Keribin wrote two articles for Tangente magazine: on clustering here and mixture models here.

10.3.2   Education

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

10.3.3   Interventions

Gilles Stoltz supervises an "atelier MATh.en.JEANS" at Lycée Douanier Rousseau and Collège Fernand Puech, Laval.

11   Scientific production

11.1   Major publications

[1]   C. Biernacki, J. Jacques and C. Keribin. 'A Survey on Model-Based Co-Clustering: High Dimension and Estimation Challenges'. In: *Journal of Classification* (17th July 2023). URL: https://hal.scie nce/hal-03769727.

[2]   E. Chzhen, C. Giraud, Z. Li and G. Stoltz. 'Small Total-Cost Constraints in Contextual Bandits with Knapsacks, with Application to Fairness'. In: *Advances in Neural Information Processing Systems*. Thirty-seventh Conference on Neural Information Processing Systems. New Orleans, United States, Dec. 2023. URL: https://hal.science/hal-04105622.

[3]   O. Coudray, C. Keribin, P. Massart and P. Pamphile. 'Risk bounds for PU learning under Selected At
       Random assumption'. In: *Journal of Machine Learning Research* 24.107 (Jan. 2023), pp. 1–31. URL:
       https://inria.hal.science/hal-03526889.

[4]   C. Giraud, Y. Issartel and N. Verzelen. 'Localization in 1D non-parametric latent space models
       from pairwise affinities'. In: *Electronic Journal of Statistics* 17.1 (1st Jan. 2023), pp. 1587–1662. DOI:
       10.1214/23-ejs2134. URL: https://hal.inrae.fr/hal-04166788.

[5]   P. Humbert, B. Le Bars, A. Bellet and S. Arlot. 'One-Shot Federated Conformal Prediction'. In: ICML
       2023 - 40th International Conference on Machine Learning. Proceedings of the 40th International
       Conference on Machine Learning (ICML). Honolulu (Hawai), United States, 23rd July 2023. URL:
       https://hal.science/hal-03981605.

11.2   Publications of the year

International journals

[6]   C. Biernacki, J. Jacques and C. Keribin. 'A Survey on Model-Based Co-Clustering: High Dimension
       and Estimation Challenges'. In: *Journal of Classification* (17th July 2023). URL: https://hal.scie
       nce/hal-03769727.

[7]   I. Bournaud and P. Pamphile. 'Une étude exploratoire sur le rôle de l'intelligence émotionnelle dans
       un dispositif d'accompagnement de primo-entrant.es à l'université'. In: *Les Annales de QPES* 2.2
       (25th Sept. 2023). URL: https://hal.science/hal-04347085.

[8]   O. Coudray, C. Keribin, P. Massart and P. Pamphile. 'Risk bounds for PU learning under Selected At
       Random assumption'. In: *Journal of Machine Learning Research* 24.107 (Jan. 2023), pp. 1–31. URL:
       https://inria.hal.science/hal-03526889.

[9]   R. Coulaud, C. Keribin and G. Stoltz. 'Modeling dwell time in a data-rich railway environment: with
       operations and passenger flows data'. In: *Transportation research. Part C, Emerging technologies*
       146 (2023), p. 103980. URL: https://hal.science/hal-03651835.

[10]   C. Giraud, Y. Issartel and N. Verzelen. 'Localization in 1D non-parametric latent space models
       from pairwise affinities'. In: *Electronic Journal of Statistics* 17.1 (1st Jan. 2023), pp. 1587–1662. DOI:
       10.1214/23-ejs2134. URL: https://hal.inrae.fr/hal-04166788.

[11]   H. Hadiji and G. Stoltz. 'Adaptation to the Range in K-Armed Bandits'. In: *Journal of Machine
       Learning Research* 24.13 (2023), pp. 1–33. URL: https://hal.science/hal-02794382.

[12]   E. Lasalle. 'Heat diffusion distance processes: a statistically founded method to analyze graph data
       sets'. In: *Journal of Applied and Computational Topology* (18th May 2023). DOI: 10.1007/s41468-
       023-00125-w. URL: https://hal.science/hal-03366848.

[13]   S. Varet, C. Lacour, P. Massart and V. Rivoirard. 'Numerical performance of penalized comparison
       to overfitting for multivariate kernel density estimation'. In: *ESAIM: Probability and Statistics* 27
       (2023), pp. 621–667. DOI: 10.1051/ps/2022018. URL: https://hal.science/hal-04131767.

International peer-reviewed conferences

[14]   A. Barrier, A. Garivier and G. Stoltz. 'On Best-Arm Identification with a Fixed Budget in Non-
       Parametric Multi-Armed Bandits'. In: ALT 2023 - The 34th International Conference on Algorithmic
       Learning Theory. Proceedings of ALT 2023. Singapour, Singapore: PMLR, Feb. 2023. URL: https:
       //hal.science/hal-03792668.

[15]   C. Biernacki, C. Boyer, G. Celeux, J. Josse, F. Laporte, M. Marbac, A. Sportisse and V. Vandewalle.
       'Impact of missing data on mixtures and clustering with illustrations in Biology and Medicine'. In:
       SPSR 2023 - The 24th annual Conference of the Romanian Society of Probability and Statistics.
       Bucarest, Romania, 21st Apr. 2023. URL: https://inria.hal.science/hal-04370870.

[16]   E. Chzhen, C. Giraud, Z. Li and G. Stoltz. 'Small Total-Cost Constraints in Contextual Bandits with
       Knapsacks, with Application to Fairness'. In: Advances in Neural Information Processing Systems.
       Vol. 36. New Orleans, United States, 2023. URL: https://hal.science/hal-04105622.

[17] R. Coulaud, C. Keribin and G. Stoltz. 'Modélisation des déplacements à bord de trains pour l'estimation de la charge à bord par zone'. In: *54es Journées de Statistique de la SFdS*. JdS 2023 - 54es Journées de Statistique de la SFdS. Bruxelles, Belgium, July 2023. URL: https://hal.science/hal-04335416.

[18] S. Gaucher, N. Schreuder and E. Chzhen. 'Fair learning with Wasserstein barycenters for non-decomposable performance measures'. In: *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*. AISTATS. Vol. 206. Valence (Espagne), Spain, 1st May 2023, pp. 2436–2459. URL: https://hal.science/hal-03767640.

[19] P. Humbert, B. Le Bars, A. Bellet and S. Arlot. 'One-Shot Federated Conformal Prediction'. In: ICML 2023 - 40th International Conference on Machine Learning. Proceedings of the 40th International Conference on Machine Learning (ICML). Honolulu (Hawai), United States, 23rd July 2023. URL: https://hal.science/hal-03981605.

[20] E. M. Saad and G. Blanchard. 'Constant regret for sequence prediction with limited advice'. In: *Proceedings of The 34th International Conference on Algorithmic Learning Theory, PMLR*. Algorithmic Learning Theory (ALT 2023). 201. Singapore, Singapore, 20th Feb. 2023, pp. 1343–1386. URL: https://hal.science/hal-03797597.

Conferences without proceedings

[21] I. Bournaud, C. Clavel, M. Gallezot, M.-J. Ramage and P. Pamphile. 'Le rapport d'étonnement comme outil pédagogique d'accompagnement pour faciliter l'acculturation des primo-entrants à l'université'. In: Symposium « Se construire en tant qu'apprenant·e au sein de l'enseignement supérieur : quels dispositifs pour favoriser un processus de trans-formation, acculturation et émancipation chez les primo-arrivant·es et adultes en reprise d'études ? ». Biennale Internationale de l'Éducation, de la Formation et des Pratiques Professionnelles. Paris, France, 20th Sept. 2023. URL: https://hal.science/hal-04363720.

Doctoral dissertations and habilitation theses

[22] E. Baroux. 'Reliability fatigue design under complex loadings : from specification to validation'. Institut Polytechnique de Paris, 19th June 2023. URL: https://theses.hal.science/tel-04257338.

[23] A. Barrier. 'Contributions to a Theory of Pure Exploration in Sequential Statistics'. Ecole normale supérieure de lyon - ENS LYON, 20th July 2023. URL: https://theses.hal.science/tel-04192097.

Reports & preprints

[24] A. Akhavan, E. Chzhen, M. Pontil and A. B. Tsybakov. *Gradient-free optimization of highly smooth functions: improved analysis and a new algorithm*. 3rd June 2023. URL: https://hal.science/hal-04275558.

[25] A. Barbier-Chebbah, C. L. Vestergaard, J.-B. Masson and E. Boursier. *Approximate information maximization for bandit games*. 6th Oct. 2023. URL: https://hal.science/hal-04246907.

[26] E. Baroux, P. Pamphile, D. Benoit, C. Andrei and R. Laurent. *Reliable fatigue design of personal vehicle chassis parts from multi-input loads and unsupervised statistical analyses*. Dec. 2023. URL: https://hal.science/hal-04375447.

[27] Z. Benomar, E. Chzhen, N. Schreuder and V. Perchet. *Addressing bias in online selection with limited budget of comparisons*. 8th Nov. 2023. URL: https://hal.science/hal-04275550.

[28] E. Chzhen, C. Giraud and G. Stoltz. *Parameter-free projected gradient descent*. 24th May 2023. URL: https://hal.science/hal-04105636.

[29] E. Chzhen and S. Schechtman. *SignSVRG: fixing signSGD via variance reduction*. 22nd May 2023. URL: https://hal.science/hal-04112556.

[30] O. Coudray, P. Bristiel, M. Dinis, C. Keribin and P. Pamphile. *Construction of fatigue criteria through Positive Unlabeled Learning.* 5th Dec. 2023. URL: https://inria.hal.science/hal-04324629.

[31] M. Jonckheere, C. Mignacco and G. Stoltz. *Symphony of experts: orchestration with adversarial insights in reinforcement learning.* 24th Oct. 2023. URL: https://hal.science/hal-04256705.

[32] P. Lacroix and M.-L. Martin. *Trade-off between predictive performance and FDR control for high-dimensional Gaussian model selection.* 27th July 2023. URL: https://hal.science/hal-03978309.

[33] P. Pamphile and I. Bournaud. *Etudier la structure latente de données multivariées à l'aide d'analyses non supervisées, en science de l'éducation.* Sept. 2023. URL: https://inria.hal.science/hal-04375594.

[34] H. Richard, E. Boursier and V. Perchet. *Constant or logarithmic regret in asynchronous multiplayer bandits.* 31st May 2023. URL: https://hal.science/hal-04273108.

[35] A. Sportisse, M. Marbac, F. Laporte, G. Celeux, C. Boyer, C. Biernacki and J. Josse. *Accompanying note : Model-based Clustering with Missing Not At Random Data.* 21st Dec. 2023. URL: https://hal.science/hal-04358192.

[36] A. Sportisse, M. Marbac, F. Laporte, G. Celeux, C. Boyer, J. Josse and C. Biernacki. *Model-based Clustering with Missing Not At Random Data.* 21st Dec. 2023. URL: https://hal.science/hal-03494674.

Other scientific publications

[37] C. Clavel, M.-J. Ramage, M. Gallezot, P. Pamphile and I. Bournaud. 'Accompagner les primo-entrants dans leur acculturation à l'université par la rédaction d'un rapport d'étonnement.' In: Journées d'Etude AIPU Section France. Université Paris-Saclay, Campus Vallée, France, 9th Nov. 2023. URL: https://hal.science/hal-04363717.