RESEARCH CENTER
Lille - Nord Europe

FIELD

Activity Report 2019

Section New Results

Edition: 2020-03-21

BONUS Project-Team

7. New Results

7.1. Decomposition-based optimization

During the year 2019, we have investigated decomposition-based optimization in the decision space as well as in the objective space. In the decision space, we have considered discrete as well as continuous problems. For discrete problems, we have reinvestigated our interval-based decomposition approach proposed in [7] as a baseline to explore ultra-scale Branch-and-Bound algorithms using Chapel [13]. The contribution is presented in Section 7.3 . For continuous problems, we have extended the geometric fractal decomposition-based approach [10] to multi-objective optimization [32] and importantly to parallel computing [17] to deal with scalability, one of the major scientific challenges as pointed out in Section 3.1 . In the objective space, we have deeply studied in [23] the surrogate-assisted multi-objective optimization based on decomposition. The contributions are summarized in the following.

7.1.1. Parallel fractal decomposition for big continuous optimization problems

Participants: El-Ghazali Talbi [contact person], Amir Nakib [Laboratoire Images, Signaux et Systèmes Intelligents (LISSI), Paris], Léo Souquet [Laboratoire Images, Signaux et Systèmes Intelligents (LISSI), Paris].

Fractal Decomposition Algorithm (FDA) is a metaheuristic that was recently proposed to solve high dimensional continuous optimization problems [10]. This approach is based on a geometric fractal decomposition which divides the search space while looking for the optimal solution. While FDA and its fractal decomposition has shown to be an effective optimization algorithm, its running time grows significantly as the problems dimension increases. To deal with this expensive computational time, a parallelized version of FDA, called Parallel Fractal Decomposition Algorithm (PFDA) is proposed in [17]. The focus is on parallelizing the exploration and exploitation phases of the original algorithm in a multi-threaded environment. The performances of PFDA are evaluated on the same Benchmark used to illustrate FDA efficiency, the SOCO 2011. It is composed of 19 functions with *dimensions going from 50 to 5000*. The results show that PFDA allows one to achieve similar performances as the original version with a significantly reduced computational time.

7.1.2. Deterministic multi-objective fractal decomposition

Participants: El-Ghazali Talbi [contact person], Amir Nakib [Laboratoire Images, Signaux et Systèmes Intelligents (LISSI), Paris], Léo Souquet [Laboratoire Images, Signaux et Systèmes Intelligents (LISSI), Paris].

We have proposed in [32] a new deterministic Multi-objective Fractal Decomposition Algorithm (Mo-FDA). The original FDA [10] was designed for single-objective large-scale continuous optimization problems. It is based on a "divide-and-conquer" strategy and a geometric fractal decomposition of the search space using hyperspheres. In this work, a scalarization approach is used to deal with multi-objective problems. The performance of Mo-FDA is compared to state-of-the-art algorithms from the literature on classical benchmarks of multi-objective optimization.

7.1.3. Surrogate-assisted multi-objective optimization based on decomposition

Participants: Nicolas Berveglieri, Bilel Derbel, Arnaud Liefooghe, Hernan Aguirre [Shinshu University, Japan], Kiyoshi Tanaka [Shinshu University, Japan].

A number of surrogate-assisted evolutionary algorithms are being developed for tackling expensive multi-objective optimization problems. On the one hand, a relatively broad range of techniques from both machine learning and multi-objective optimization can be combined for this purpose. Different taxonomies exist in order to better delimit the design choices, advantages and drawbacks of existing approaches. On the other hand, assessing the relative performance of a given approach is a difficult task, since it depends on the characteristics of the problem at hand. In [23], we focus on surrogate-assisted approaches using objective space decomposition as a core component. We propose a refined and fine-grained classification, ranging from EGO-like approaches to filtering or pre-screening. More importantly, we provide a comprehensive comparative study of a representative selection of state-of-the-art methods, together with simple baseline algorithms. We rely on selected benchmark functions taken from the bbob-biobj benchmarking test suite, that provides a variable range of objective function difficulties. Our empirical analysis highlights the effect of the available budget on the relative performance of each approach, and the impact of the training set and of the machine learning model construction on both solution quality and runtime efficiency.

7.2. ML-assisted optimization

As pointed out in our research program 3.2 , we investigate the ML-assisted optimization following two directions: building efficiently surrogates to deal with expensive black-box objective functions and automatically building and predicting/improving the performance of metaheuristics through landscape/problem structure analysis. Regarding surrogate-assisted optimization, we put the focus in [31] on mixed discrete-continuous optmization problems, one of the major challenges of this topic. In addition, we focused on the evolution control and batch parallelism to deal with another challenge which consists in efficiently integrating surrogate models (Bayesian neural networks in this case) in evolutionary algorithms [FGCS-Briffoteaux, à compléter]. From the application point of view, we applied our approaches to three different real-world simulation-based problems in the context of three collaborations: multi-stage optimal scheduling of virtual power plants (collaboration with electrical engineering department of UMONS University, Belgium) in [27], aerospace vehicle design (collaboration with ONERA, Paris) in [31], and resource allocation for the Tuberculosis edidemic control (Monash University) in [FGCS-Briffoteax, à compléter]. The two latter ones are summarized in the following sections. Regarding the second direction, we thoroughly study in [16] the impact of landscape characteristics on the performance of search heuristics for black-box multi-objective combinatorial optimization problems. We also introduce in [26] new insightful features for continuous exploratory landscape analysis and algorithm selection. In addition, as pointed out in 3.2 , variable selection is highly important to deal with BOPs. In [28], in collaboration with our partners from Shinshu University we come out with an efficient method to classify variables influencing convergence and increase their recombination rate. The contributions are summarized in the following.

7.2.1. Surrogate-assisted optimization of constrained mixed variable problems: application to the design of a launch vehiclethrust frame.

Participants: El-Ghazali Talbi [contact person], Julien Pelamatti, Loïc Brevault [ONERA], Mathieu Balesdent [ONERA], Yannick Guerin [CNES].

Within the framework of complex systems design, such as launch vehicles, numerical optimization is an essential tool as it allows to reduce the design process time and costs. The inclusion of discrete variables in the design process allows to extend the applicability of numerical optimization methods to a broader number of systems and sub-systems. In [31], a recently proposed adaptation of the Efficient Global Optimization method (EGO) for constrained mixed-variable problems is applied to the design optimization of a launch vehicle thrust frame, which depends on both continuous sizing parameters and discrete variables characterizing the number of structural reinforcements. The EGO adaptation that is considered is based on a redefinition of the Gaussian Process kernel as a product between a standard continuous kernel and a second kernel representing the covariance between the discrete variable values. From the results obtained on an analytical test-case as well as on the launch vehicle thrust frame design optimization, it is shown that the use of the mixed-variable EGO algorithm allows to converge towards the neighborhoods of the problems optima with fewer function evaluations when compared to reference optimization algorithms.

7.2.2. Parallel Batched Bayesian Neural Network-assisted GA versus q-EGO

Participants: Guillaume Briffoteaux, Maxime Gobert, Jan Gmys, Nouredine Melab [contact person], Romain Ragonnet [School of Public Health and Preventive Medicine, Monash University, Australia], Mohand Mezmaz [University of Mons, Blegium], Daniel Tuyttens [University of Mons, Blegium].

Surrogate-based optimization has been widely used to deal with expensive black-box simulation-based objective functions. The use of a surrogate model allows to reduce the number of calls to the costly simulator. In (SWEVO, Briffoteaux et al., 2020), the Efficient Global Optimization (EGO) reference framework is challenged by a Bayesian Neural Network-assisted Genetic Algorithm, namely BNN-GA. The Bayesian Neural Network (BNN) surrogate provides an uncertainty measure of the prediction that allows to compute the Expected Improvement of a candidate solution in order to improve the exploration of the objective space. BNN is also more reliable than Kriging models for high-dimensional problems and faster to set up thanks to its incremental training. In addition, we propose a batch-based approach for the parallelization of BNN-GA that is challenged by a parallel version of EGO, called q-EGO. Parallel computing is a complementary way to deal with the computational burden of simulation-based optimization. The comparison of the two parallel approaches is experimentally performed through several benchmark functions and two real-world problems within the scope of Tuberculosis Transmission Control (TBTC). The results demonstrate the efficiency and scalability of the parallel batched BNN-GA for small budgets, outperforming it for larger budgets on the benchmark testbed. A significant improvement of the solutions is obtained for two TBTC problems. Finally, our study proves that parallel batched BNN-GA is a viable alternative to q-EGO approaches being more suitable for high-dimensional problems and parallelization impact.

7.2.3. Landscape-aware performance prediction and algorithm selection for single- and multi-objective evolutionary optimization

Participants: Arnaud Liefooghe, Bilel Derbel, Fabio Daolio [ASOS.com, UK], Sébastien Verel [LISIC, Université du Littoral Côte d'Opale], Hernan Aguirre [Shinshu University, Japan], Kiyoshi Tanaka [Shinshu University, Japan].

Extracting a priori knowledge informing about the landscape underlying an unknown optimization problem has been proved extremely useful for different purposes, such as designing finely-tuned algorithms, predicting algorithm performance and designing automated portfolio-based solving approaches. Considering black-box continuous single-objective optimization problems, in [26], we adopt an exploratory landscape analysis approach providing a unified methodology for integrating landscape features into sophisticated machine learning techniques. More precisely, we consider the design of novel informative and cheap landscape features on the basis of the search tree constructed by the so-called SOO (Simultaneous Optimistic Optimization) algorithm, which is arguably a global optimizer coming from the machine learning community and having its foundations in the multi-armed bandit theory. We thereby provide empirical evidence on the accuracy of the proposed features for both predicting high-level problem properties, and tackling the algorithm selection problem with respect to a given portfolio of available solvers and a broad range of optimisation problems taking from the specialized literature and exposing different degrees of difficulty.

Considering black-box combinatorial multi-objective optimization problems, in [16], we expose and contrast the impact of landscape characteristics on the performance of search heuristics for black-box multi-objective combinatorial optimization problems. A sound and concise summary of features characterizing the structure of an arbitrary problem instance is identified and related to the expected performance of global and local dominance-based multi-objective optimization algorithms. We provide a critical review of existing features tailored to multi-objective combinatorial optimization problems, and we propose additional ones that do not require any global knowledge from the landscape, making them suitable for large-size problem instances. Their intercorrelation and their association with algorithm performance are also analyzed. This allows us to assess the individual and the joint effect of problem features on algorithm performance, and to highlight the main difficulties encountered by such search heuristics. By providing effective tools for multi-objective landscape analysis, we highlight that multiple features are required to capture problem difficulty, and we provide further insights into the importance of ruggedness and multimodality to characterize multi-objective combinatorial landscapes.

7.2.4. Estimating Relevance of Variables for Effective Recombination

Participants: Arnaud Liefooghe, Bilel Derbel, Sébastien Verel [LISIC, Université du Littoral Côte d'Opale], Taishi Ito [Shinshu University, Japan], Hernan Aguirre [Shinshu University, Japan], Kiyoshi Tanaka [Shinshu University, Japan].

Dominance and its extensions, decomposition, and indicator functions are well-known approaches used to design MOEAs. Algorithms based on these approaches have mostly sought to enhance parent selection and survival selection. In addition, several variation operators have been developed for MOEAs. In [28], we focus on the classification and selection of variables to improve the effectiveness of solution search. We propose a method to classify variables that influence convergence and increase their recombination rate, aiming to improve convergence of the approximation found by the algorithm. We incorporate the proposed method into NSGA-II and study its effectiveness using three-objective DTLZ and WFG benchmark functions, including unimodal, multimodal, separable, non-separable, unbiased, and biased functions. We also test the effectiveness of the proposed method on a real-world bi-objective problem. Simulation results verify that the proposed method can contribute to achieving faster and better convergence in several kinds of problems, including the real-world problem.

7.3. Towards ultra-scale Big Optimization

During the year 2019, we have addressed the ultra-scale optimization research line following two main directions: designing efficient optimization algorithms dealing with scalability in terms of number of processing cores and/or in terms of the size of tackled problem instances. For the first direction, the challenge is to take into account in addition to the traditional performance objective the productivity awareness which is highly important to deal with the increasing complexity of medern supercomputers. With the short-term perspective of establishing a collaboration with one of the major HPC builders (Cray Inc.), we have investigated in [13], [24], [25] the exascale-aware Chapel language. Regarding the second direction, we contributed to solve difficult unsolved flow-shop [15] and Quadratic Assignement Problem (QAP) [22] permutation problem instances using moderate-scale parallelism. The contributions are summarized in the following.

7.3.1. Towards ultra-scale Branch-and-Bound using a high-productivity language

Participants: Tiago Carneiro Pessoa, Jan Gmys, Nouredine Melab, Daniel Tuyttens [University of Mons, Blegium].

Productivity is crucial for designing ultra-scale algorithms able to harness modern supercomputers which are increasingly complex, including millions of processing cores and heterogeneous building-block devices. In [13], we investigate the partitioned global address space (PGAS)-based approach using Chapel for the productivity-aware design and implementation of distributed Branc-and-Bound (B&B) for solving large optmization problems. The proposed algorithms are intensively experimented using the Flow-shop scheduling problem as a test-case. The Chapel-based implementation is compared to its MPI+X-based traditionally used counterpart in terms of performance, scalability, and productivity. The results show that Chapel is much more expressive and up to $7.8\times$ more productive than MPI+Pthreads. In addition, the Chapel-based search presents performance equivalent to MPI+Pthreads for its best results on 1024 cores and reaches up to 84% of the linear speedup. However, there are cases where the built-in load balancing provided by Chapel cannot produce regular load among computer nodes. In such cases, the MPI-based search can be up to $4.2\times$ faster and reaches speedups up to $3\times$ higher than its Chapel-based counterpart. Thorough feedback on the experience is given, pointing out the strengths and limitations of the two opposite approaches (Chapel vs. MPI+X). To the best of our knowledge, the present study is pioneering within the context of exact parallel optimization.

7.3.2. A computationally efficient Branch-and-Bound algorithm for the permutation flow-shop scheduling problem

Participants: Jan Gmys, Nouredine Melab, Mohand Mezmaz [University of Mons, Blegium], Daniel Tuyttens [University of Mons, Blegium].

In [15], we propose an efficient Branch-and-Bound (B&B) algorithm for the permutation flow-shop problem (PFSP) with makespan objective. We present a new node decomposition scheme that combines dynamic branching and lower bound refinement strategies in a computationally efficient way. To alleviate the computational burden of the two-machine bound used in the refinement stage, we propose an online learning-inspired mechanism to predict promising couples of bottleneck machines. The algorithm offers multiple choices for branching and bounding operators and can explore the search tree either sequentially or in parallel on multi-core CPUs. In order to empirically determine the most efficient combination of these components, a series of computational experiments with 600 benchmark instances is performed. A main insight is that the problem size, as well as interactions between branching and bounding operators substantially modify the trade-off between the computational requirements of a lower bound and the achieved tree size reduction. Moreover, we demonstrate that parallel tree search is a key ingredient for the resolution of large problem instances, a strong super-linear speedups can be observed. An overall evaluation using two well-known benchmarks indicates that the proposed approach is superior to previously published B&B algorithms. For the first benchmark we report the exact resolution – within less than 20 minutes – of two instances defined by 500 jobs and 20 machines that remained open for more than 25 years, and for the second a total of 89 improved best-known upper bounds, including proofs of optimality for 74 of them.

7.3.3. A Parallel Tabu Search for the Large-scale Quadratic Assignment Problem

Participants: Omar Abdelkafi, Bilel Derbel, Arnaud Liefooghe.

Parallelization is an important paradigm for solving massive optimization problems. Understanding how to fully benefit form the aggregated computing power and what makes a parallel strategy successful is a difficult issue. In [22], we propose a simple parallel iterative tabu search (PITS) and study its effectiveness with respect to different experimental settings. Using the quadratic assignment problem (QAP) as a case study, we first consider different small-and medium-size instances from the literature and then tackle a large-size instance that was rarely considered due the its inherent solving difficulty. In particular, we show that a balance between the number of function evaluations each parallel process is allowed to perform before resuming the search is a critical issue to obtain an improved quality.

DEFROST Project-Team

7. New Results

7.1. Soft robots locomotion and manipulation control using FEM simulation and quadratic programming

In this work, we proposed a method to control the motion of soft robots able to manipulate objects or roll from one place to another. We used the Finite Element Method (FEM) to simulate the deformations of the soft robot, its actuators, and its environment. To find the inverse model of the robot interacting with obstacles, and with constraints on its actuators, we wrote the problem as a quadratic program with complementarity constraints. The novelty of this work was that friction contacts (sticking contact only) is taken into account in the optimization process, allowing the control of these specific tasks that are locomotion and manipulation. We proposed a formulation that simplifies the optimization problem, together with a dedicated solver [22]. The algorithm had real-time performance and handles evolving environments as long as we know them. To show the effectiveness of the method, we presented several numerical examples, and a demonstration on a real robot (see Figure 2 and 3).

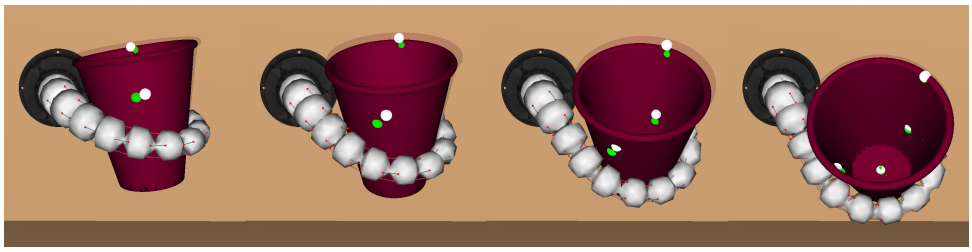


Figure 2. Simulation of a soft gripper holding a deformable cup subject to gravity. Here we optimize the cables displacements to control the position/orientation of the cup. A phantom of the cup target is shown in transparency. The cup have four controlled points represented by the green spheres. The corresponding targets are represented by the white spheres.

Figure 3. Real soft robot actuated online using the output of the simulation. In this scenario, using our control framework, we are able to control the orientation of the real plastic cup (see the attached video).

7.2. Toward Shape Optimization of Soft Robots

This year, we obtained new results on shape optimization for soft robotics where the shape is optimized for a given soft robot usage. To obtain a parametric optimization with a reduced number of parameters, we relied on an approach where the designer progressively refines the parameter space and the fitness function until a satisfactory design is obtained. In our approach, we automatically generate FEM simulations of the soft robot and its environment to evaluate a fitness function while checking the consistency of the solution. Finally, we have coupled our framework to an evolutionary optimization algorithm, and demonstrated its use for optimizing the design of a deformable leg of a locomotive robot. A paper presenting the approach was accepted at IEEE/International Conference On Soft-Robotics2019 [29].

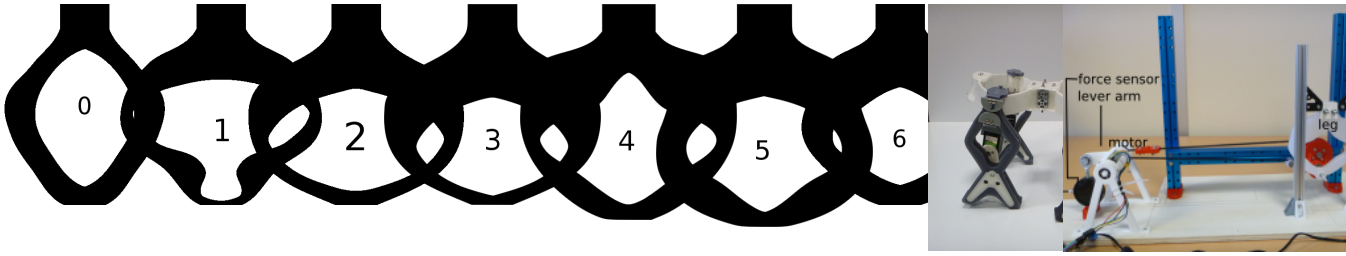


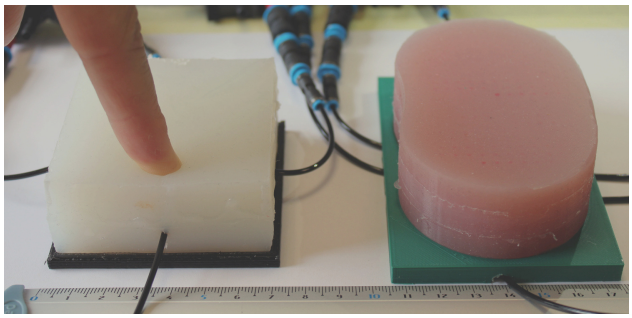
Figure 4. Initial shape (labelled 0) for the optimization and different shapes (labelled 1 to 6) obtained by numerical experiments. Pictures represents the robot with the legs and the test best that has been used to verify the accuracy of the simulation

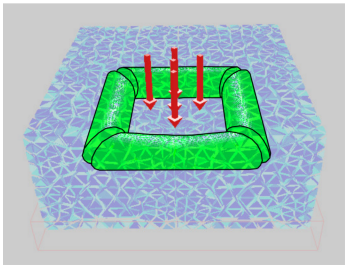
7.3. Modeling Novel Soft Mechanosensors based on Air-Flow Measurements

In this work, we introduce a new pneumatic mechanosensor dedicated to Soft Robotics and propose a generic method to reconstruct the magnitude of a contact-force acting on it. This is illustrated by Fig. 5 . Changes in cavity volumes inside a soft silicon pad are measured by air-flow sensors. The resulting mechanosensor is characterized by its high sensitivity, repeatability, dynamic range and accurate localization capability in 2D. Using a regression found by machine learning techniques we can predict the contact location and force magnitude accurately when the force magnitudes are within the range of the training data. To be able to provide a more general model, a novel approach based on a Finite Element Method (FEM) is introduced. We formulate an optimization problem, which yields the contact load that best explains the observed changes in cavity volumes. This method makes no assumptions on the force range, the shape of the soft pad or the shape of its cavities. The prediction of the force also results in a model for the deformation of the soft pad. We characterize our sensor and evaluate two designs, a soft pad and a kidney-shaped sensor, in different scenarios. A paper was accepted for the journal *Robotics and Automation Letters (RA-L)* [5]. Furthermore, an extended abstract was accepted at the *RoboTac 2019 Workshop* at *IROS 2019*, leading to a presentation of a demo of the proposed technology.

7.4. Calibration and External Force Sensing for Soft Robots using an RGB-D Camera

Benefiting from the deformability of soft robots, calibration and force sensing for soft robots are possible using an external vision-based system, instead of embedded mechatronic force sensors. In this work, we first propose a calibration method to calibrate both the sensor-robot coordinate system and the actuator inputs. This task is addressed through a sequential optimization problem for both variables. We also introduce an external force





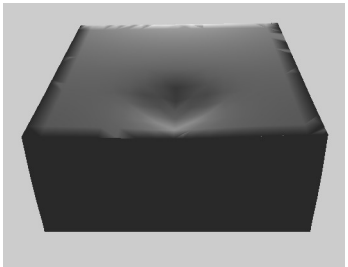


Figure 5. In this work, we show two designs of a novel soft mechanosensor made out of silicone (top, a soft pad and a kidney). When an external force is applied, the volume of cavities embedded in the silicone changes (left). This change in volume is registered through air-flow sensors. Using machine learning and FEM-based techniques, we show that it is possible to estimate the location and magnitude of an external force on the mechanosensor. Using the FEM also yields an estimation of the deformation of the sensor (left and right).

sensing system based on a real-time Finite Element (FE) model with the assumption of static configurations, and which consists of two steps: force location detection and force intensity computation. The algorithm that estimates force location relies on the segmentation of the point cloud acquired by an RGB-D camera. Then, the force intensities can be computed by solving an inverse quasi-static problem based on matching the FE model with the point cloud of the soft robot. As for validation, the proposed strategies for calibration and force sensing have been tested using a parallel soft robot driven by four cables (see figure 6 and reference [18]).

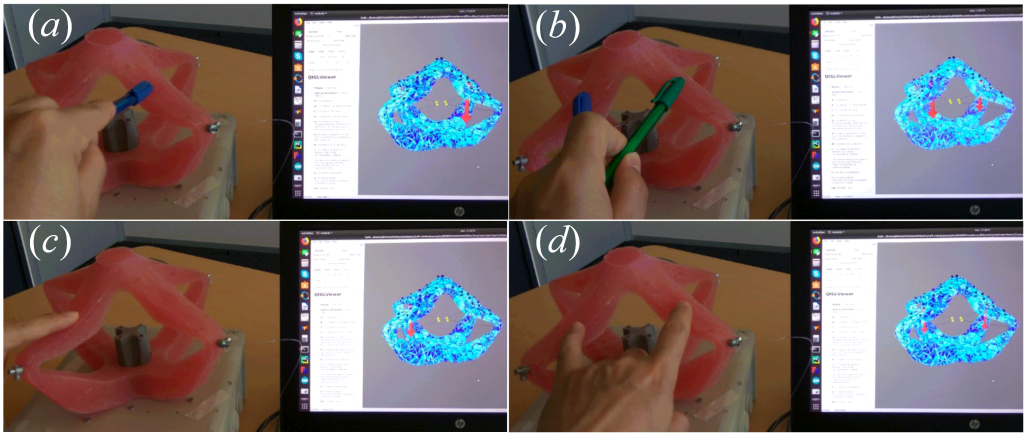


Figure 6. Screenshot of external force sensing. The robot has four cables with constant length for the experiment. (a) and (c) show one external force on the actuated soft robot. (b) and (d) show case with two external forces.

7.5. Motion Control of Cable-Driven Continuum Catheter Robot through Contacts

Catheter-based intervention plays an important role in minimally invasive surgery. For the closed-loop control of catheter robot through contacts, the loss of contact sensing along the entire catheter might result in task failure. To deal with this problem, we propose a decoupled motion control strategy which allows to control insertion and bending independently. We model the catheter robot and the contacts using the Finite Element Method. Then, we combine the simulated system and the real system for the closed-loop motion control. The control inputs are computed by solving a quadratic programming (QP) problem with a linear complementarity problem (LCP). A simplified method is proposed to solve this optimization problem by converting it into a standard QP problem. Using the proposed strategy, not only the control inputs but also the contact forces along the entire catheter can be computed without using force sensors. Finally, we validate the proposed methods using both simulation and experiments on a cable-driven continuum catheter robot for the real-time motion control through contacts [17].

7.6. Control Design for Soft Robots based on Reduced Order Model

Inspired by nature, soft robots promise disruptive advances in robotics. Soft robots are naturally compliant and exhibit nonlinear behavior, which makes their study challenging. No unified framework exists to control these robots, especially when considering their dynamics. This work proposes a methodology to study this type of robots around a stable equilibrium point. It can make the robot converge faster and with reduced oscillations to a desired equilibrium state. Using computational mechanics, a large-scale dynamic model of the robot is obtained and model reduction algorithms enable the design of low order controller and observer. A real robot is used to demonstrate the interest of the results [14].

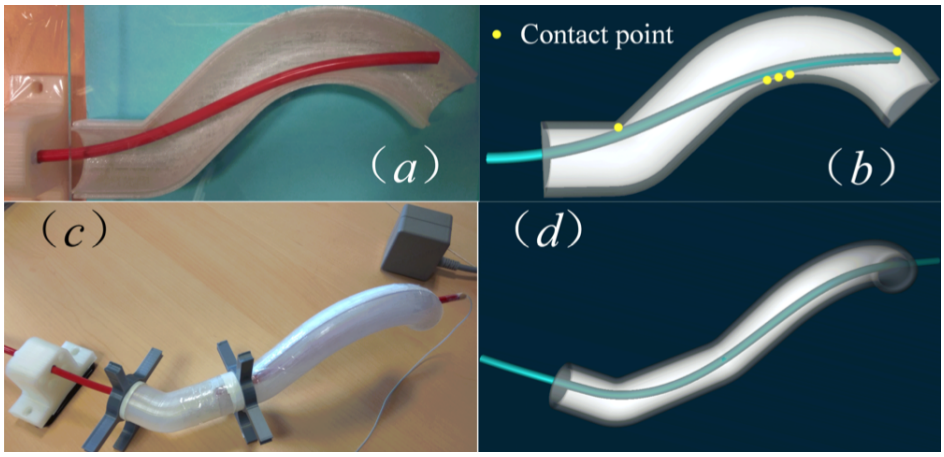


Figure 7. (a) and (c) present experimental setups for the validation. (b) and (d) show that we take into account the contact and provide accurate results in the simulation

FUN Project-Team

6. New Results

6.1. Security and Verification

Participants: Rehan Malak, Allan Blanchard, Antoine Gallais, Valeria Loscri, Nathalie Mitton.

6.1.1. Security

Numerous medium access control (MAC) have been proposed for Low-power Lossy Networks (LLNs) over the recent years. They aim at ensuring both energy efficiency and robustness of the communication transmissions. Nowadays, we observe deployments of LLNs for potentially critical application scenarios (e.g., plant monitoring, building automation), which require both determinism and security guarantees. They involve battery-powered devices which communicate over lossy wireless links. Radio interfaces are turned off by a node as soon as no traffic is to be sent or relayed. Denial-of-sleep attacks consist in exhausting the devices by forcing them to keep their radio on. In [21], we focus on jamming attacks whose impact can be mitigated by approaches such as time-division and channel hopping techniques. We use the IEEE 802.15.4e standard to show that such approaches manage to be resistant to jamming but yet remain vulnerable to selective jamming. We discuss the potential impacts of such onslaughts, depending on the knowledge gained by the attacker, and to what extent envisioned protections may allow jamming attacks to be handled at upper layers.

6.1.2. Verification

Modern verification projects continue to offer new challenges for formal verification. One of them is the linked list module of Contiki, a popular open-source operating system for the Internet of Things. It has a rich API and uses a particular list representation that make it different from the classical linked list implementations. Being widely used in the OS, the list module is critical for reliability and security. A recent work verified the list module using ghost arrays. In [17], [35], we report on a new verification effort for this module. Realized in the Frama-C/Wp tool, the new approach relies on logic lists. A logic list provides a convenient high-level view of the linked list. The specifications of all functions are now proved faster and almost all automatically, only a small number of auxiliary lemmas and a couple of assertions being proved interactively in Coq. The proposed specifications are validated by proving a few client functions manipulating lists. During the verification, a more efficient implementation for one function was found and verified. We compare the new approach with the previous effort based on ghost arrays, and discuss the benefits and drawbacks of both techniques.

While deductive verification is increasingly used on real-life code, making it fully automatic remains difficult. The development of powerful SMT solvers has improved the situation, but some proofs still require interactive theorem provers in order to achieve full formal verification. Auto-active verification relies on additional guiding annotations (assertions, ghost code, lemma functions, etc.) and provides an important step towards a greater automation of the proof. However, the support of this methodology often remains partial and depends on the verification tool. [18] presents an experience report on a complete functional verification of several C programs from the literature and real-life code using auto-active verification with the C software analysis platform Frama-C and its deductive verification plugin . The goal is to use automatic solvers to verify properties that are classically verified with interactive provers. Based on our experience, we discuss the benefits of this methodology and the current limitations of the tool, as well as proposals of new features to overcome them.

6.2. Visible Light Communication

Participants: Antonio Costanzo, Valeria Loscri.

Visible Light Communication (VLC) exploits optical frequencies, diffused by usual LED lamps, for adding data communication features to illuminating systems. This paradigm has attracted a growing interest in both scientific and industrial community in the latter decade. Nevertheless, classical wireless communication mechanisms for physical and Medium Access Control (MAC) layers are hardly available for VLC, due to the massive external interference caused by sunlight. Moreover, effects related to the data frames features need to be taken into account in order to improve the effectiveness of the VLC paradigm. Such as an instance, the preamble length of a packet in order to synchronize the data transmission represents an important factor in VLC. A too long preamble allows a better synchronization while impacting negatively in terms of overhead. Nevertheless, a too short preamble may be not effective for synchronizing the transmission. The more suitable selection of the preamble length is strictly related to the noise environments. In order to make an adaptive selection able to choice the more suitable preamble length, we have designed and integrated in our VLC system a machine learning algorithm based on multi-arm bandit approach, in order to dynamically select the best configuration [28]. Another important approach to face the high interference impacting on the VLC performance is represented by the treatment of the noise through a signal processing approach in order to estimate it and proceed with a mitigation of the noise [10], [29], [36], [30]. This approach has been implemented and tested by the means of a real prototype. Results obtained show the effectiveness of a similar approach.

6.3. Alternative communications

Participant: Valeria Loscri.

In the last few years, there has been an increasing interest in the study of "alternative communication" paradigms, ranging from the exploitation of the visible light as carrier information, the exploitation of a different portion of the spectrum in the THz frequency, the leverage of artificial molecules for transmitting information (i.e. artificial molecular communication). Another interesting approach is consisting on a different perspective of the interaction between the signals and the environment. Right now, the environment has always been considered as something that cannot be "changed", a kind of obstacle for the wireless transmission. A new paradigm is arising, based on the metamaterial surface, where the interaction between the signals and the environment can be adapted in order to improve the performance of the communication.

6.3.1. TeraHez communications

In [32] and [23] we have investigated a metasurface and how the design of this metasurface has to be realized in order to adapt the behavior of the optical and THz signals based on the specific application considered.

6.3.2. Molecular communications

Concerning the artificial molecular communication paradigm, a fundamental aspect to be considered is that the most of times the target application of this type of communication is a biological system. A fundamental question arising is then: how maximize the effectiveness of the communication by keeping a low impact in terms of interference for the biological system? We have tried to answer to these fundamental questions by considering a signal processing approach in [11], [19].

6.4. Long range communications

Participants: Nathalie Mitton, Brandon Foubert, Ibrahim Amadou.

In the context of smart farming, communications still pose a key challenge. Ubiquitous access to the Internet is not available worldwide, and battery capacity is still a limitation. Inria and the Sencrop company are collaborating to develop an innovative solution for wireless weather stations, based on multi-technology communications, to enable smart weather stations deployment everywhere around the globe. We discuss this model in [12] and assess the quality of a LoRA signal in different conditions [16].

6.5. Vehicular networks

Participants: Nathalie Mitton, Valeria Loscri.

6.5.1. Positioning

Typical Global Navigation Satellite System (GNSS) receivers offer precision in the order of meters. This error margin is excessive for vehicular safety applications, such as forward collision warning, autonomous intersection management, or hard braking sensing. In [14] we develop CooPS, a GNSS positioning system that uses Vehicle to Vehicle (V2V) and Vehicle to Infrastructure (V2I) communications to cooperatively determine absolute and relative position of the ego-vehicle with enough precision. To that end, we use differential GNSS through position vector differencing to acquire track and across-track axes projections, employing elliptical and spherical geometries. We evaluate CooPS performance by carrying out real experiments using off-the-shelf IEEE 802.11p equipment at the campus of the Federal University of Rio de Janeiro. We obtain an accuracy level under 1.0 and 1.5 m for track (where-in-lane) and across-track (which-lane) axes, respectively. These accuracy levels were achieved using a 2.5 m accuracy circular error probable (CEP) of 50% and a 5 Hz navigation update rate GNSS receiver.

6.5.2. Vehicular social networks

In recent years, the concept of social networking combined with the Internet of Vehicles has brought to the definition of the Social IoV (SIoV) paradigm, i.e., a social network where every vehicle is capable of establishing social relationships in an autonomous way with other vehicles or road infrastructure equipment. In SIoV, social networking is applied to vehicular networks according to how social ties are built upon, i.e., either among vehicles or humans. An analysis of the SIoT-based social relations in a vehicular network scenario for establishing a Social Internet of Vehicles and providing insights on this growing research area [33]. By considering the specific features of the Online Social Networks (ONSs) and Vehicular Social Networks (VSNs), we realize that there are limitations and advantages on both these systems. In [15] we have proposed SOVER, a hybrid OSN-VSN framework, allowing the communication between both the communities, the OSNs and VSNs. In [24] we investigate the twofold nature of SIoV, both based on human factors and relationships and as an instance of the Social Internet of Things (SIoT. Based on this twofold nature, it is possible to distinguish different applications and use-cases.)

6.6. On the use of controlled mobility

Participant: Nathalie Mitton.

Relying on controlled mobility as enabled by drones or robots could be a great asset for task management, data collection or quality of network deployment.

6.6.1. Robots

Robots and controlled mobility can help in the dynamic coverage of an area. In [22], we address the problem of defining a wireless sensor network by deploying sensors with the aim of guaranteeing the coverage of the area and the connectivity among the sensors. The wireless sensor networks are widely studied since they provide several services, e.g., environmental monitoring and target tracking. We consider several typologies of sensors characterized by different sensing and connectivity ranges. A cost is associated with each typology of sensors. In particular, the higher the sensing and connectivity ranges, the higher the cost. We formulate the problem of deploying sensors at minimum cost such that each sensor is connected to a base station with either a one-or a multi-hop and the area is full covered. We present preliminary computational results by solving the proposed mathematical model, on several instances. We provide a simulation-based analysis of the performances of such a deployment from the routing perspective.

Robots could be helpful when called upon an alert sent by sensors. But to intervene quickly, they need to locate or follow back the alert source as fast as possible. Two new algorithms (GFGF1 and GFGF2) for event finding in wireless sensor and robot networks based on the Greedy-Face-Greedy (GFG) routing are proposed in [27]. The purpose of finding the event (reported by sensors) is to allocate the task to the closest robot to act upon the event. Using two scenarios (event in or out of the network) and two topologies (random and random with hole) it is shown that GFGF1 always find the closest robot to the event but with more than twice higher communication cost compared to GFG, especially for the outside of the network scenario. GFGF2 features

more than 4 times communication cost reduction compared to GFG but with percentage of finding the closest robot up to 90%.

6.6.2. Drones

Disaster scenarios are particularly devastating in urban environments, which are generally very densely populated. Disasters not only endanger the life of people, but also affect the existing communication infrastructure. In fact, such an infrastructure could be completely destroyed or damaged; even when it continues working, it suffers from high access demand to its resources within a short period of time, thereby compromising the efficiency of rescue operations. [31], [25] leverage the ubiquitous presence of wireless devices (e.g., smartphones) in urban scenarios to assist search and rescue activities following a disaster. This work considers multi-interface wireless devices and drones to collect emergency messages in areas affected by natural disasters. Specifically, it proposes a collaborative data collection protocol that organizes wireless devices in multiple tiers by targeting a fair energy consumption in the whole network, thereby extending the network lifetime. Moreover, it introduces a scheme to control the path of drones so as to collect data in a short time. Simulation results in realistic settings show that the proposed solution balances the energy consumption in the network by means of efficient drone routes, thereby effectively assisting search and rescue operations.

6.7. Self-organization, routing and orchestration

Participants: Nathalie Mitton, Valeria Loscri, Brandon Foubert.

By offering low-latency and context-aware services, fog computing will have a peculiar role in the deployment of Internet of Things (IoT) applications for smart environments. Unlike the conventional remote cloud, for which consolidated architectures and deployment options exist, many design and implementation aspects remain open when considering the latest fog computing paradigm. In [9], we focus on the problems of dynamically discovering the processing and storage resources distributed among fog nodes and, accordingly, orchestrating them for the provisioning of IoT services for smart environments. In particular, we show how these functionalities can be effectively supported by the revolutionary Named Data Networking (NDN) paradigm. Originally conceived to support named content delivery, NDN can be extended to request and provide named computation services, with NDN nodes acting as both content routers and in-network service executors. To substantiate our analysis, we present an NDN fog computing framework with focus on a smart campus scenario, where the execution of IoT services is dynamically orchestrated and performed by NDN nodes in a distributed fashion. A simulation campaign in ndnSIM, the reference network simulator of the NDN research community, is also presented to assess the performance of our proposal against state-of-the-art solutions. Results confirm the superiority of the proposal in terms of service provisioning time, paid at the expenses of a slightly higher amount of traffic exchanged among fog nodes.

[26] proposes FLY-COPE, a complete self-organization architecture that relies on cooperative communications and drone-assisted data collection, allowing a fast location of victims and rescuing operation organization in disaster relief operation. FLY-COPE mainly combines two components: i) a ground component that spontaneously emerges from any communicating devices (piece of infrastructure, mobile phone, etc) that cooperate to alert rescuers and remain all alive as long as possible and ii) an aerial component comprising UAV to communicate efficiently with ground devices. We show by simulation and/or by experimentation that each component of FLY-COPE allows substantial energy saving for efficient and fast disaster response.

The IPv6 Routing Protocol for Low-Power and Lossy Networks (RPL) builds a Direction Oriented Directed Acyclic Graph (DODAG) rooted at one node. This node may act as a border router to provide Internet connectivity to the members of the DODAG but such a situation creates a single point of failure. Upon border router failure, all nodes connected to the DODAG are affected as all ongoing communications are instantly broken and no new communications can be initiated. Moreover, nodes close to the border router should forward traffic from farther nodes in addition to their own, which may cause congestion and energy depletion inequality. In [20], we specify a full solution to enable border router redundancy in RPL networks. To achieve this, we propose a mechanism leveraging cooperation between colocated RPL networks. It enables failover to maintain Internet connectivity and load balancing to improve the overall energy consumption and

bandwidth. Our contribution has been implemented in Contiki OS and was evaluated through experiments performed on the FIT IoT-LAB testbed.

INOCS Project-Team

7. New Results

7.1. Large scale complex structure optimization

Joint order batching and picker routing problem: Order picking is the process of retrieving products from inventory. It is mostly done manually by dedicated employees called pickers and is considered the most expensive of warehouse operations. To reduce the picking cost, customer orders can be grouped into batches that are then collected by traveling the shortest possible distance. We proposed an industrial case study for the HappyChic company where the warehouse has an acyclic layout: pickers are not allowed to backtrack. We developed a two-phase heuristic approach to solve this industrial case [64]. Moreover, we propose an exponential linear programming formulation to tackle the joint order batching and picker routing problem. Variables, or columns, are related to the picking routes in the warehouse. Computing such routes is generally an intractable routing problem and relates to the well known traveling salesman problem (TSP). Nonetheless, the rectangular warehouse's layouts can be used to efficiently solve the corresponding TSP and take into account in the development of an efficient subroutine, called oracle. We therefore investigate whether such an oracle allows for an effective exponential formulation. Experimented on a publicly available benchmark, the algorithm proves to be very effective. It improves many of the best known solutions and provides very strong lower bounds. Finally, this approach is also applied to the HappyChic industrial case to demonstrate its interest for this field of application [67], [45].

Logistics network design problem: Planning transportation operations within a supply chain is a difficult task that is often outsourced to logistics providers, in practice. At the tactical level, the problem of distributing products through a multi-echelon network is defined in the literature as the Logistics Service Network Design Problem (LSNDP). We study a LSNDP variant inspired by the management of restaurant supply chains. In this problem, a third party carrier seeks to cost-effectively source and fulfill customer demands of products through a tri-echelon supply chain composed of suppliers, warehouses, and customers. We propose an exact solution method based on partial Benders decompositions, where the master problem is strengthened by the addition of aggregated information derived from the subproblem. More specifically, we introduce a high-level dynamic Benders approach where the aggregated information used to strengthen the master is refined iteratively. In an extensive computational study, we demonstrate that our dynamic Benders strategy produces provably high-quality solutions and we validate the interest of refining the master problem in the course of a partial Benders decomposition-based scheme [72].

Multi commodity vehicle routing problem: We study vehicle routing problems considering multiple commodities, with applications in the local fresh food supply chains. The studied supply chain contains two echelons with three sets of actors: suppliers, distribution centers and customers. Suppliers are farmers that produce some fresh foods. Distribution centers are in charge of consolidation and delivery of the products to customers. Distribution centers collect products from the suppliers that perform direct trips. Products are delivered to the customers with a fleet of vehicles performing routes. Each customer requires several commodities, and the farmers produce a limited quantity of these commodities. For the minimization of the transportation cost, it is beneficial that a single customer is delivered by several vehicles. However, for the convenience of the customer, it is imposed that a single commodity is delivered at once by a single vehicle. Hence, different commodities have been explicitly considered. The complete problem is named Multi-Commodity two-echelon Distribution Problem (MC2DP). The restricted problem that addresses only the delivery from a single distribution center is named Commodity constrained Split Delivery Vehicle Routing Problem (C-SDVRP). We first propose a heuristic based on the Adaptive Large Neighborhood Search (ALNS) for the C-SDVRP [22]. Then, we address the whole problem (MC2DP) with collection and delivery operations and multiple distribution centers. In order to tackle this complex problem, we propose to decompose the problem: collection and delivery are sequentially solved [53], [54].

Generalized routing problems: We study routing problems that arise in the context of last mile delivery when multiple delivery options are proposed to the customers. The most common option to deliver packages is home/workplace delivery. Besides, the delivery can be made to pick-up points such as dedicated lockers or stores. In recent years, a new concept called trunk/in-car delivery has been proposed. Here, customers' packages can be delivered to the trunks of cars. Our goal is to model and develop efficient solution approaches for routing problems in this context, in which each customer can have multiple shipping locations. First, we study the single-vehicle case in the considered context, which is modeled as a Generalized Traveling Salesman Problem with Time Windows (GTSPTW). Four mixed integer linear programming formulations and an efficient branch-and-cut algorithm are proposed. Then, we study the multi-vehicle case which is denoted Generalized Vehicle Routing Problem with Time Windows (GVRPTW). An efficient column generation based heuristic is proposed to solve it [60], [61], [59], [37].

Joint management of demand and offer for last-mile delivery systems: E-commerce is a thriving market around the world and suits very well the busy lifestyle of today's customers. This growing in e-commerce poses a huge challenge for transportation companies, especially in the last mile delivery. We addressed first a fleet composition problem for last-mile delivery service. This problem occurs at a tactical level when the composition of the fleet has to be decided in advance. It is the case for companies that offer last-mile delivery service. Most of them subcontract the transportation part to local carriers and have to decide the day before which vehicles will be needed to cover a partially known demand. We assumed that the distribution area is divided into a limited number of delivery zones and the time horizon into time-slots. The demand is characterized by packages to be transported from pick-up zones to delivery zones given a delivery time slot. We have proposed an optimization problem aiming jointly to manage the offer and the demand. More precisely, discrete choice models representing the choices made by couriers and customers are integrated into the same optimization model. The originality of the contribution is based on the integration of variables of different natures in the same model and the development of integrated resolution methods. On the basis of a closed form of discrete choice models, we have reformulated the problem as a non-linear optimization problem. The resolution of this model by classical solvers requires coupling exact methods with heuristics in order to define a first initial solution [47], [58].

Delay Management in Public Transportation: The Delay Management Problem arises in Public Transportation networks, and is characterized by the necessity of connections between different vehicles. The attractiveness of Public Transportation networks is strongly related to the reliability of connections, which can be missed when delays or other unpredictable events occur. Given a single initial delay at one node of the network, the Delay Management Problem is to determine which vehicles have to wait for the delayed ones, with the aim of minimizing the dissatisfaction of the passengers. We derived strengthened mixed integer linear programming formulations and new families of valid inequalities for that problem. The implementation of branch-and-cut methods and tests on a benchmark of instances taken from real networks show the potential of the proposed formulations and cuts [20].

Discrete Ordered Median Problem: The discrete ordered median problem consists in locating p facilities in order to minimize an ordered weighted sum of distances between clients and closest open facility. We formulate this problem as a set partitioning problem using an exponential number of variables. Each variable corresponds to a set of demand points allocated to the same facility with the information of the sorting position of their corresponding costs. We develop a column generation approach to solve the continuous relaxation of this model. Then, we apply a branch-price-and-cut algorithm to solve small to large sized instances of DOMP in competitive computational time [21].

Genome wide association studies: We studied the Polymorphic Alu Insertion Recognition Problem (PAIRP). Alu (Arthrobacter luteus) forms a major component of repetitive DNA and are frequently encountered during the genotyping of individuals. The basic approach to find Alus consists of (1) aligning sequence reads from a set of individual(s) with respect to a reference genome and (2) comparing the possible Alu insertion induced by the alignment with the Alu insertions positions already known for the reference genome. The sequence genome of the reference individual is known and will be highly similar, but not identical, to the genome of the individual(s) being sequenced. Hence, at some locations they will diverge. Some of this divergence is due to

the insertion of Alu polymorphisms. Detecting Alus has a central role in the field of Genetic Wide Association Studies because basic elements are a common source of mutation in humans. We investigated the PAIRP relationship with the the Clique Partitioning of Interval Graphs (CPIG). Our results [29] provide insights of the complexity of the problem, a characterization of its combinatorial structure and an exact approach based on Integer Linear Programming to exactly solve the correspond instances.

A branch-and-cut algorithm for the maximum k-balanced subgraph of a signed graph: A signed graph is k-balanced if its vertex set can be partitioned into at most k sets in such a way that positive edges are found only within the sets and negative edges go between sets. The maximum k-balanced subgraph problem is the problem of finding a subgraph of a graph G that is k-balanced and maximum according to the number of vertices. This problem has applications in clustering problems appearing in collaborative vs conflicting environments. We provide a representatives formulation for the problem and present a partial description of the associated polytope, including the introduction of strengthening families of valid inequalities. A branch-and-cut algorithm is described for finding an optimal solution to the problem. An ILS metaheuristic is implemented for providing primal bounds for this exact method and a branching rule strategy is proposed for the representatives formulation. Computational experiments, carried out over a set of random instances and on a set of instances from an application, show the effectiveness of the valid inequalities and strategies adopted in this work [75].

Feature Selection in Support Vector Machine: This work focuses on support vector machine (SVM) with feature selection. A MILP formulation is proposed for the problem. The choice of suitable features to construct the separating hyperplanes has been modelled in this formulation by including a budget constraint that sets in advance a limit on the number of features to be used in the classification process. We propose both an exact and a heuristic procedure to solve this formulation in an efficient way. Finally, the validation of the model is done by checking it with some well-known data sets and comparing it with classical classification methods [24].

7.2. Bilevel Programming

Pricing for Energy Management: Power systems face higher flexibility requirements from generation to consumption due to the increasing penetration of non-controllable distributed renewable energy. In this context, demand side management aims at reducing excessive load fluctuation and match the price of energy to their real cost for the grid. Pricing models for demand side management methods are traditionally used to control electricity demand. First, we proposed bilevel pricing models to explore the relationship between energy suppliers and customers who are connected to a smart grid. The smart grid technology allows customers to keep track of hourly prices and shift their demand accordingly, and allows the provider to observe the actual demand response to its pricing strategy. Moreover, we assumed that the smart grid optimizes the usage of a renewable energy generation source and a storage capacity. Results over a rolling horizon were obtained (Léonard Von Niederhausern PhD thesis [76]). Next, we considered four types of actors: furnishers sell electricity, local agents trade and consume energy, aggregators trade energy and provide energy to end-users, who consume it. This gives rise to three levels of optimization. The interaction between aggregators and their end-users is modeled with a bilevel program, and so is the interaction between furnishers, and local agents and aggregators. Since solving bilevel programs is difficult in itself, solving trilevel programs requires particular care. We proposed three possible approaches, two of them relying on a characterization of the intermediary optimization level [13], [76].

Finally, Time and-Level-of-Use is a recently proposed energy pricing scheme, designed for the residential sector and providing suppliers with robust guarantee on the consumption. We formulate the supplier decision as a bilevel, bi-objective problem optimizing for both financial loss and guarantee. A decomposition method is proposed, related to the optimal value transformation. It allows for the computation of an exact solution by finding possible Pareto optimal candidate solutions and then eliminating dominated ones. Numerical results on experimental residential power consumption data show the method effectively finds the optimal candidate solutions while optimizing costs only or incorporating risk aversion at the lower-level [38], [46].

Linear bilevel optimization: One of the most frequently used approaches to solve linear bilevel optimization problems consists in replacing the lower-level problem with its Karush–Kuhn–Tucker (KKT) conditions and by reformulating the KKT complementarity conditions using techniques from mixed-integer linear optimization. The latter step requires to determine some big-M constant in order to bound the lower level's dual feasible set such that no bilevel-optimal solution is cut off. In practice, heuristics are often used to find a big-M although it is known that these approaches may fail. In [69], we consider the hardness of two proxies for the above mentioned concept of a bilevel-correct big-M. First, we prove that verifying that a given big-M does not cut off any feasible vertex of the lower level's dual polyhedron cannot be done in polynomial time unless P = NP. Second, we show that verifying that a given big-M does not cut off any optimal point of the lower level's dual problem (for any point in the projection of the high-point relaxation onto the leader's decision space) is as hard as solving the original bilevel problem.

Market regulation: We proposed a bilevel programming model to study a problem of market regulation through government intervention. One of the main characteristics of the problem is that the government monopolizes the raw material in one industry, and competes in another industry with private firms for the production of commodities. Under this scheme, the government controls a state-owned firm to balance the market; that is, to minimize the difference between the produced and demanded commodities. On the other hand, a regulatory organism that coordinates private firms aims to maximize the total profit by deciding the amount of raw material bought from the state-owned firm. Two equivalent single-level reformulations are proposed to solve the problem. Additionally, three heuristic algorithms are designed to obtain good-quality solutions with low computational effort. Extensive computational experimentation is carried out to measure the efficiency of the proposed solution methodologies. A case study based on the Mexican petrochemical industry is presented. Additional instances generated from the case study are considered to validate the robustness of the proposed heuristic algorithms [28].

Product pricing: One of the main concerns in management and economic planning is to sell the right product to the right customer for the right price. Companies in retail and manufacturing employ pricing strategies to maximize their revenues. The Rank Pricing Problem considers a unit-demand model with unlimited supply and uniform budgets in which customers have a rank-buying behavior. Under these assumptions, the problem is first analyzed from the perspective of bilevel pricing models and formulated as a non linear bilevel program with multiple independent followers. We also present a direct non linear single level formulation. Two different linearizations of the models are carried out and two families of valid inequalities are obtained which, embedded in the formulations by implementing a branch-and-cut algorithm, allow us to tighten the upper bound given by the linear relaxation of the models. We show the efficiency of the formulations, the branch-and-cut algorithms and some preprocessing through extensive computational experiments [73].

Next in [68], we analyze a product pricing problem with single-minded customers, each interested in buying a bundle of products. The objective is to maximize the total revenue and we assume that supply is unlimited for all products. We contribute to a missing piece of literature by giving some mathematical formulations for this single-minded bundle pricing problem. We first present a mixed-integer nonlinear program with bilinear terms in the objective function and the constraints. By applying classical linearization techniques, we obtain two different mixed-integer linear programs. We then study the polyhedral structure of the linear formulations and obtain valid inequalities based on an RLT-like framework. We develop a Benders decomposition to project strong cuts from the tightest model onto the lighter models. We conclude this work with extensive numerical experiments to assess the quality of the mixed-integer linear formulations, as well as the performance of the cutting plane algorithms and the impact of the preprocessing on computation times.

Bilevel Minimum Spanning Tree Problem: Consider a graph G whose edge set is partitioned into a set of red edges and a set of blue edges, and assume that red edges are weighted and contain a spanning tree of G. Then, the Bilevel Minimum Spanning Tree Problem (BMSTP) consists in pricing (i.e., weighting) the blue edges in such a way that the total weight of the blue edges selected in a minimum spanning tree of the resulting graph is maximized. We propose different mathematical formulations for the BMSTP based on the properties of the Minimum Spanning Tree Problem and the bilevel optimization. We establish a theoretical and empirical

comparison between these new formulations and we also provide reinforcements that together with a proper formulation are able to solve medium to big size instances [26].

Bilevel programming models for location problems: First, we addressed a multi-product location problem in which a retail firm has several malls with a known location. A particular product comes in p types. Each mall has a limited capacity for products to be sold at that location, so the firm has to choose what products to sold at what mall. Furthermore, the firm can apply discrete levels of discount on the products. The objective of the firm is to find what products to sell at which mall, with what level of discount, so that its profit is maximized. Consumers are located in points of the region. Each consumer has a different set of acceptable products, and will purchase one of these, or none if it is not convenient for her. Consumers maximize their utility. The agents (firm and consumers) play a Stackelberg game, in which the firm is the leader and the customers the follower. Once the firm decides the products to sell at each mall and the possible discounts, consumers purchase (or not) one of their acceptable products wherever their utility is maximized. We model the problem using bilevel formulations, which are compared on known instances from the literature [74]. Second we studied a location problem of controversial facilities. On the one hand, a leader chooses among a number of fixed potential locations which ones to establish. On the second hand, one or several followers who, once the leader location facilities have been set, choose their location points in a continuous framework. The leader's goal is to maximize some proxy to the weighted distance to the follower's location points, while the follower(s) aim is to locate his location points as close as possible to the leader ones. We develop the bilevel location model for one follower and for any polyhedral distance, and we extend it for several followers and any so-called p-norm. We prove the NP-hardness of the problem and propose different mixed integer linear programming formulations. Moreover, we develop alternative Benders decomposition algorithms for the problem. Finally, we report some computational results comparing the formulations and the Benders decompositions on a set of instances [23].

Stackelberg games: First we analyzed general Stackelberg games (SGs) and Stackelberg security games (SSGs). SGs are hierarchical adversarial games where players select actions or strategies to optimize their payoffs in a sequential manner. SSGs are a type of SGs that arise in security applications, where the strategies of the player that acts first consist in protecting subsets of targets and the strategies of the followers consist in attacking one of the targets. We review existing mixed integer optimization formulations in both the general and the security setting and present new formulations for the the second one. We compare the SG formulations and the SSG formulations both from a theoretical and a computational point of view. We identify which formulations provide tighter linear relaxations and show that the strongest formulation for the security version is ideal in the case of one single attacker. Our computational experiments show that the new formulations can be solved in shorter times [18].

Second, we formulate a Stackelberg Security game that coordinates resources in a border patrol problem. In this security domain, resources from different precincts have to be paired to conduct patrols in the border due to logistic constraints. Given this structure the set of pure defender strategies is of exponential size. We describe the set of mixed strategies using a polynomial number of variables but exponentially many constraints that come from the matching polytope. We then include this description in a mixed integer formulation to compute the Strong Stackelberg Equilibrium efficiently with a branch and cut scheme. Since the optimal patrol solution is a probability distribution over the set of exponential size, we also introduce an efficient sampling method that can be used to deploy the security resources every shift. Our computational results evaluate the efficiency of the branch and cut scheme developed and the accuracy of the sampling method. We show the applicability of the methodology by solving a real world border patrol problem [15].

Third, in [39], we discuss the impact of fairness constraints in Stackelberg Security Games. Fairness constraints can be used to avoid discrimination at the moment of implementing police patrolling. We present two ways of modelling fairness constraints, one with a detailed description of the population and the other with labels. We discuss the implementability of these constraints. In the case that the constraints are not implementable we present models to retrieve pure strategies in a way that they are the closest in average to the set of fairness constraints.

Finally, in [65], we focus on Stackelberg equilibria for discounted stochastic games. We begin by formalizing the concept of Stationary Strong Stackelberg Equlibrium (SSSE) policies for such games. We provide classes

of games where the SSSE exists, and we prove via counterexamples that SSSE does not exist in the general case. We define suitable dynamic programming operators whose fixed points are referred to as Fixed Point Equilibrium (FPE). We show that the FPE and SSSE coincide for a class of games with Myopic Follower Strategy. We provide numerical examples that shed light on the relationship between SSSE and FPE and the behavior of Value Iteration, Policy Iteration and Mathematical programming formulations for this problem. Finally, we present a security application to illustrate the solution concepts and the efficiency of the algorithms studied.

7.3. Robust/Stochastic programming

Locating stations in a one-way electric car sharing system under demand uncertainty: In [16], we focused on a problem of locating recharging stations in one-way station based electric car sharing systems which operate under demand uncertainty. We modeled this problem as a mixed integer stochastic program and develop a Benders decomposition algorithm based on this formulation. We integrated a stabilization procedure to our algorithm and conduct a large-scale experimental study on our methods. To conduct the computational experiments, we developed a demand forecasting method allowing to generate many demand scenarios. The method was applied to real data from Manhattan taxi trips.

Bookings in the European Gas Market: Characterisation of Feasibility and Computational Complexity Results: As a consequence of the liberalisation of the European gas market in the last decades, gas trading and transport have been decoupled. At the core of this decoupling are so-called bookings and nominations. Bookings are special long-term capacity right contracts that guarantee that a specified amount of gas can be supplied or withdrawn at certain entry or exit nodes of the network. These supplies and withdrawals are nominated at the day-ahead. These bookings then need to be feasible, i.e., every nomination that complies with the given bookings can be transported. While checking the feasibility of a nomination can typically be done by solving a mixed-integer nonlinear feasibility problem, the verification of feasibility of a set of bookings is much harder. We consider the question of how to verify the feasibility of given bookings for a number of special cases. For our physics model we impose a steady-state potential-based flow model and disregard controllable network elements. We derive a characterisation of feasible bookings, which is then used to show that the problem is in coNP for the general case but can be solved in polynomial time for linear potential-based flow models. Moreover, we present a dynamic programming approach for deciding the feasibility of a booking in tree-shaped networks even for nonlinear flow models [25]. Further, in [71], we show that the feasibility of a booking also can be decided in polynomial time on single-cycle networks.

Robust bilevel programs: Bilevel optimization problems embed the optimality conditions of a sub-problem into the constraints of a decision-making process. A general question of bilevel optimization occurs where the lower-level is solved (only) to near-optimality. Solving bilevel problems under limited deviations of the lower-level variables was introduced under the term "ϵ-approximation" of the pessimistic bilevel problem. In [77] the authors define special properties and a solution method for this variant in the so-called independent case, i.e., where the lower-level feasible set is independent of the upper-level decision. In [66], we generalized the approach of *Wiesemann et al. 2013*, to problems with constraints involving upper- and lower-level variables in the constraints at both levels. The purpose of this generalization is to protect the upper-level feasibility against uncertainty of near-optimal solutions of the lower-level. We call this near-optimal robustness and the generalization is a near-optimal robust bilevel problem (NORBiP). NORBiP is a bilinear bilevel problem, and this makes it very hard in general. We have defined and implemented a solution algorithm for the linear linear NORBiP [66].

LINKS Project-Team

7. New Results

7.1. Querying Heterogeneous Linked Data

7.1.1. Data Integration and Schema Validation

Data integration requires knowledge about the structure of the various data. Such a structure is usually described by schemas. While for relational databases, schemas are hard-coded, this is not the case for many other formats. In XML for instance, several schema formalisms exists, such as DTD, XML Schema or Schematron. The Links Project-Team investigate the problem of defining schemas and use them to data, in particular for RDF and JSON Formats.

With P. Wieczorek of the University of Wroclaw, Poland, S. Staworko et al. have studied the containment problem of ShEx schemas for RDF documents in *PODS* [10].

Also, J. Dusart develops under the supervision of I. Boneva and S. Staworko the software *ShEx Validator* so as to foster the practical usage of ShEx. It is also worth noting that ShEx is now being adopted by several institutions such as *WikiData*.

7.1.2. Aggregates

Aggregation refers to computations that are alien to mere logical data manipulation (e.g. such as in relational algebra). Typically, aggregation means counting the number of answers, or performing other kinds of statistics. We have a slightly larger understanding as we may also include enumerating all answers with a *small delay*. Aggregation algorithms are generally subtle as they in most cases avoid the explicit generation of the whole set of answers. We study aggregation problems within the ANR project *Aggreg* coordinated by Niehren.

In the same spirit, Capelli et al. (in a joint work with Mengel from the CNRS in Lens) showed at *STACS* [7] a new knowledge compilation procedure which allows a polynomial algorithm to test the satisfiability quantified Boolean formulas with bounded tree width. In *Theory of Computing Systems*, [25], Capelli also gave a taxonomy of results according to various restrictions of tree-width of graphs.

Also, in *Theory of Computing Systems*, [25], Capelli gave a taxonomy of results according to various restrictions of tree-width of graphs.

Finally, in an article in *JCSS* [14], F. Capelli (with Bergougnoux and Kanté from Bordeaux and Clermont-Ferrand) propose an algorithm for counting the number of transversals (i.e. subset of nodes intersecting all hyperedges) in some hypergraphs.

7.1.3. Certain Query Answering

When data is incomplete, logical constraints and knowledge about its intended structure help to infer the answers of queries. This inference problem is known as *certain query answering*.

L. Gallois and S. Tison [6] presented in *IJCAI* - one of the main conferences of Artificial Intelligence. L. Gallois and S. Tison study boundedness of the chase procedure in the context of positive existential rules, providing decidability results for several classes and outlining the complexity of the problem. This work is done in collaboration with P. Bourhis and Graphik team-project. These results also belong to the PhD thesis of L. Gallois [11] supervised by S. Tison and P. Bourhis.

7.2. Managing Dynamic Linked Data

7.2.1. Complex Event Processing

Complex event processing requires to answer queries on streams of complex events, i.e., nested words or equivalently linearizations of data trees, but also to produce dynamically evolving data structures as output.

In an article published in *LATA* [17], I. Boneva, J. Niehren and M. Sakho studied certain query answering for hyperstreams - which are collections of connected streams - with *complex events* (i.e. that correspond to tree patterns). They showed that the problem is EXP-complete in general, and obtained PTIME algorithms when restricted to *linear* tree patterns (possibly with compression) and to deterministic tree automata.

LOKI Project-Team

7. New Results

7.1. Introduction

According to our research program, in the next two to five years, we will study dynamics of interaction along three levels depending on interaction time scale and related user's perception and behavior: *Micro-dynamics*, *Meso-dynamics*, and *Macro-dynamics*. Considering phenomena that occur at each of these levels as well as their relationships will help us to acquire the necessary knowledge (Empowering Tools) and technological bricks (Interaction Machine) to reconcile the way interactive systems are designed and engineered with human abilities. Our strategy is to investigate issues and address challenges for all of the three levels of dynamics. Last year we focused on micro-dynamics since it concerns very fundamental knowledge about interaction and relates to very low-level parts of interactive systems. In 2019 we were able to build upon those results (micro), but also to enlarge the scope of our studies within larger interaction time scales, especially at the meso-dynamic level. Some of these results have also contributed to our objective of defining the basic principles of an Interaction Machine.

7.2. Micro-dynamics

Participants: Géry Casiez [contact person], Sylvain Malacria, Mathieu Nancel, Thomas Pietrzak.

7.2.1. Latency & Transfer functions

End-to-end latency in interactive systems is detrimental to performance and usability, and comes from a combination of hardware and software delays. While these delays are steadily addressed by hardware and software improvements, it is at a decelerating pace. In parallel, short-term input prediction has recently shown promising results to compensate for latency, in both research and industry.

In the context of the collaborative TurboTouch project, we proposed a method based on a frequency-domain approximation of a non-causal ideal predictor with a finite impulse response filter. Given a sufficiently rich dataset, the parameters of the filter can be either optimized off-line or tuned on-line with the proposed adaptive algorithm. The performance of the proposed solution is evaluated in an experimental study consisting of drawings on a touchscreen [13].

On the related topic of transfer functions, we proposed a switched dynamic model to model indirect pointing tasks with a computer mouse. The model contains a ballistic movement phase governed by a nonlinear model in Lurie form and a corrective movement phase described by a linear visual-feedback system. The stability of the model was evaluated and the derived model was then validated with experimental data acquired in a pointing task with a mouse. Numerical comparison to pointing models available in the literature is also provided [12].

7.2.2. 3D interaction

Raycasting is the most common target pointing technique in virtual reality environments. However, performance on small and distant targets is impacted by the accuracy of the pointing device and the user's motor skills. Current pointing facilitation techniques are currently only applied in the context of the virtual hand, i.e. for targets within reach. We proposed enhancements to Raycasting: filtering the ray, and adding a controllable cursor on the ray to select the nearest target (Figure 2). We ran a series of studies for the design of the visual feedforward, filtering technique, as well as a comparative study between different 3D pointing techniques. Our results show that highlighting the nearest target is one of the most efficient visual feedforward technique. We also show that filtering the ray reduces error rate in a drastic way. Finally we show the benefits of RayCursor compared to Raycasting and another technique from the literature [19], [14].

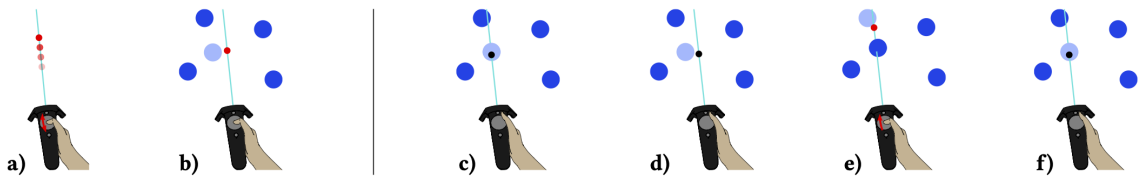


Figure 2. Illustration of manual RayCursor: a) the user controls a cursor along the ray using relative displacements of their thumb on the controller's touchpad; b) the target closest to the cursor is highlighted. Illustration of semi-auto RayCursor: c) by default, it works like Raycasting. The cursor (in black) is positioned at the intersection with a target; d) the target remains selected if the cursor moves out of the target, until it is closer to another target; e) the user can manually move the cursor using the controller's touchpad, to select another target (the cursor turns red to indicate manual mode); f) if the user does not touch the touchpad for 1s, the cursor returns to its behaviour described in c).

7.3. Meso-dynamics

Participants: Axel Antoine, Marc Baloup, Géry Casiez, Stéphane Huot, Edward Lank, Sylvain Malacria, Mathieu Nancel, Thomas Pietrzak [contact person], Thibault Raffaillac, Marcelo Wanderley.

7.3.1. Production of illustrative supports

Trace figures are contour drawings of people and objects that capture the essence of scenes without the visual noise of photos or other visual representations. Their focus and clarity make them ideal representations to illustrate designs or interaction techniques. In practice, creating those figures is a tedious task requiring advanced skills, even when creating the figures by tracing outlines based on photos. To mediate the process of creating trace figures, we introduce the open-source tool Esquisse (Figure 3). Informed by our taxonomy of 124 trace figures, Esquisse provides an innovative 3D model staging workflow, with specific interaction techniques that facilitate 3D staging through kinematic manipulation, anchor points and posture tracking. Our rendering algorithm (including stroboscopic rendering effects) creates vector-based trace figures of 3D scenes. We validated Esquisse with an experiment where participants created trace figures illustrating interaction techniques, and results show that participants quickly managed to use and appropriate the tool [18].

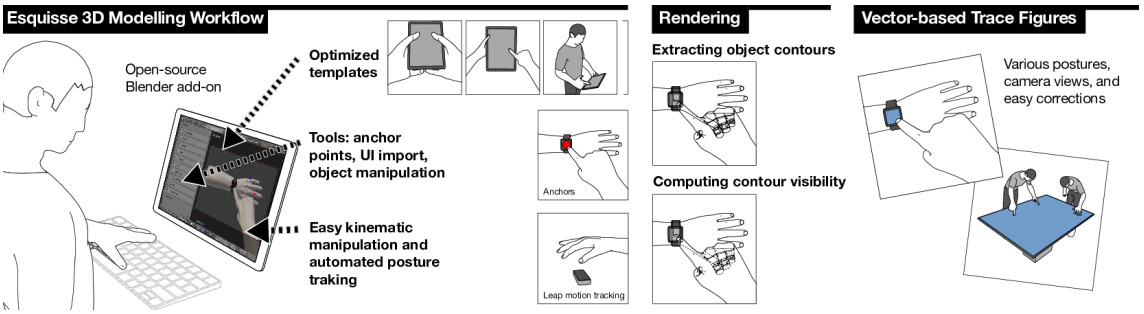


Figure 3. Workflow of Esquisse: (left) facilitating staging of 3D scenes with templates, anchor points, direct UI import, kinematic manipulation and automated posture tracking, (center) Esquisse's algorithm rendering the tracing images, and (right) final vector-based trace figures.

7.3.2. Impact of confirmation modes on expert interaction techniques adoption

Expert interaction techniques such as gestures or keyboard shortcuts are more efficient than traditional WIMP techniques because it is often faster to recall a command than to navigate to it. However, many users seem to be reluctant to switch to expert interaction. We hypothesized the cause might be the aversion to making errors. To test this, we designed two intermediate modes for the FastTap interaction technique, allowing quick confirmation of what the user has retrieved from memory, and quick adjustment if she made an error. We investigated the impact of these modes and of various error costs in a controlled study, and found that participants adopted the intermediate modes, that these modes reduced error rate when the cost of errors was high, and that they did not substantially change selection times. However, while it validates the design of our intermediate modes, we found no evidence of greater switch to memory-based interaction, suggesting that reducing error rate is not sufficient to motivate the adoption of expert use of techniques [25].

7.3.3. Effect of the context on mobile interaction

7.3.3.1. Pointing techniques for eyewear using a simulated pedestrian environment

Eyewear displays allow users to interact with virtual content displayed over real-world vision, in active situations like standing and walking. Pointing techniques for eyewear displays have been proposed, but their social acceptability, efficiency, and situation awareness remain to be assessed. Using a novel street-walking simulator, we conducted an empirical study of target acquisition while standing and walking under different levels of street crowdedness. Results showed that indirect touch was the most efficient and socially acceptable technique, and that in-air pointing was inefficient when walking. Interestingly, the eyewear displays did not improve situation awareness compared to the control condition [23].

7.3.3.2. Studying smartphone motion gestures in private or public contexts

We also investigated the effect of social exposure on smartphone motion gestures. We conducted a study where participants performed sets of motion gestures on a smartphone in both private and public locations. Using data from the smartphone's accelerometer, we found that the location had a significant effect on both the duration and intensity of the participants' gestures. We concluded that it may not be sufficient for gesture input systems to be designed and calibrated purely in private lab settings. Instead, motion gesture input systems for smartphones may need to be aware of the changing context of the device and to account for this in algorithms that interpret gestural input [26].

7.4. Macro-dynamics

Participants: Stéphane Huot, Sylvain Malacria [contact person], Nicole Pong.

7.4.1. Awareness, usage and discovery of hidden controls

Revealing a hidden widget with a dedicated sliding gesture is a common interaction design in today's handheld devices. Such "Swhidgets" (for swipe-revealed hidden widgets) provide a fast (and sometime unique) access to some commands. Interestingly, swhidgets do not follow conventional design guidelines in that they have no explicit signifiers, and users have to discover their existence before being able to use them. We conducted the first two studies specifically targeted to this type of interface design, investigating iOS users' experience with swhidgets. The first study conducted in a laboratory setting investigated which Swhidgets are spontaneously used by participants when prompted to perform certain operations on an iOS device. The second study conducted via an online survey platform, investigated which Swhidgets users reported to know and use. Combined, our studies provide the following main insights on awareness, usage and discovery of Swhidgets by middle-aged and technology-friendly users. Our results suggest that Swhidgets are moderately but unevenly known by participants, yet the awareness and the discovery issues of this design is worthy of further discussion [21].

7.5. Interaction Machine

Two of our contributions this year relate specifically to our Interaction Machine project.

7.5.1. Definition of Brain-Computer Interfaces

Regardless of the term used to designate them, Brain-Computer Interfaces are “Interfaces” between a user and a computer in the broad sense of the term. We provided a perspective to discuss how BCIs have been defined in the literature from the day the term was introduced by Jacques Vidal. From a Human-Computer Interaction perspective, we propose a new definition of Brain-Computer Interfaces as “any artificial system that transforms brain activity into input of a computer process” [24]. As they are interfaces, their definition should not include the finality and objective of the system they are used to interact with. To illustrate this, we compared BCIs with other widely used Human-Computer Interfaces, and draw analogies in their conception and purpose. This definition would help better encompassing for such interfaces in systems design, and more generally inform on how to better manage diverse forms of input in an Interaction Machine.

7.5.2. Software architecture for interactive systems

On the software engineering side, we have proposed a new Graphical User Interface (GUI) and Interaction framework based on the Entity-Component-System model (ECS) [22]. In this model, interactive elements (Entities) are characterized only by their data (Components). Behaviors are managed by continuously running processes (Systems) which select entities by the Components they possess. This model facilitates the handling of behaviors and promotes their reuse. It provides developers with a simple yet powerful composition pattern to build new interactive elements with Components. It materializes interaction devices as Entities and interaction techniques as a sequence of Systems operating on them. We have implemented these principles in the Polyphony toolkit in order to experiment the ECS model in the context of GUIs programming. It has proven to be useful and efficient for modeling standard interaction techniques, and we are now exploring its benefits for prototyping and implementing more advanced methods in a modular way. It also raises some interesting challenges about performance and scalability that we will explore further.

7.5.3. From the dynamics of interaction to an Interaction Machine

Several of our new results this year also informed our global objective of building an Interaction Machine. At the micro-dynamics level, as last year, our work on prediction algorithms and transfer functions highlighted the need for accessing low-level input data and to have flexible input management to be able to reliably predict current finger position and compensate for latency. Our work on new selection methods in 3D also highlighted the importance of easing the combination of input events from multiple sources and of data filtering to achieve better interaction. As it also leverages the real time aspect of the perception-action coupling for efficient interaction, it also confirms the need for efficient and low-latency input management stacks. These results give us the first leads to redefine input management and input events propagation in order to better account for human factors in interactive systems, and to extend the possibilities for designing more efficient and expressive interaction methods.

At the meso-dynamics level, our studies on the adoption of expert interaction techniques and of the impact of the context in performing interaction gestures highlighted the need for both adaptable and adaptive systems (e.g. context-based calibration of gesture recognition algorithms), which require more modular and flexible system architectures in order to enable real-time parametrization or even switching interaction techniques. These results also resonate with those at the micro-dynamics level, since they suggest strong links between users’ behaviors and strategies (meso) and their low-level perception mechanisms (micro) that should be better taken into account in the design of interactive systems.

These conclusions and observations will be the basis for our investigations on the topic next year. We will in particular focus on the redefinition of the input stack and on applying the ECS model to the whole architecture of an interactive system.

MAGNET Project-Team

7. New Results

7.1. Natural Language Processing

Multi-Lingual Dependency Parsing

In [1], MATHIEU DEHOUCK presents his work on Word Representation and Joint Training for Syntactic Analysis. Syntactic analysis is a key step in working with natural languages. With the advances in supervised machine learning, modern parsers have reached human performances. However, despite the intensive efforts of the dependency parsing community, the number of languages for which data have been annotated is still below the hundred, and only a handful of languages have more than ten thousands annotated sentences. In order to alleviate the lack of training data and to make dependency parsing available for more languages, previous research has proposed methods for sharing syntactic information across languages. By transferring models and/or annotations or by jointly learning to parse several languages at once, one can capitalise on languages grammatical similarities in order to improve their parsing capabilities. However, while words are a key source of information for mono-lingual parsers, they are much harder to use in multi-lingual settings because they vary heavily even between very close languages. Morphological features on the contrary, are much more stable across related languages than word forms and they also directly encode syntactic information. Furthermore, it is arguably easier to annotate data with morphological information than with complete dependency structures. With the increasing availability of morphologically annotated data using the same annotation scheme for many languages, it becomes possible to use morphological information to bridge the gap between languages in multi-lingual dependency parsing.

In his thesis, MATHIEU DEHOUCK has proposed several new approaches for sharing information across languages. These approaches have in common that they rely on morphology as the adequate representation level for sharing information. We therefore also introduce a new method to analyse the role of morphology in dependency parsing relying on a new measure of morpho-syntactic complexity. The first method uses morphological information from several languages to learn delexicalised word representations that can then be used as feature and improve mono-lingual parser performances as a kind of distant supervision. The second method uses morphology as a common representation space for sharing information during the joint training of model parameters for many languages. The training process is guided by the evolutionary tree of the various language families in order to share information between languages historically related that might share common grammatical traits. We empirically compare this new training method to independently trained models using data from the Universal Dependencies project and show that it greatly helps languages with few resources but that it is also beneficial for better resourced languages when their family tree is well populated. We eventually investigate the intrinsic worth of morphological information in dependency parsing. Indeed not all languages use morphology as extensively and while some use morphology to mark syntactic relations (via cases and persons) other mostly encode semantic information (such as tense or gender). To this end, we introduce a new measure of morpho-syntactic complexity that measures the syntactic content of morphology in a given corpus as a function of preferential head attachment. We show through experiments that this new measure can tease morpho-syntactic languages and morpho-semantic languages apart and that it is more predictive of parsing results than more traditional morphological complexity measures.

Modal sense classification with task-specific context embeddings Sense disambiguation of modal constructions is a crucial part of natural language understanding. Framed as a supervised learning task, this problem heavily depends on an adequate feature representation of the modal verb context. Inspired by recent work on general word sense disambiguation, we propose in [8] a simple approach of modal sense classification in which standard shallow features are enhanced with task-specific context embedding features. Comprehensive experiments show that these enriched contextual representations fed into a simple SVM model lead to significant classification gains over shallow feature sets.

Learning Rich Event Representations and Interactions for Temporal Relation Classification Most existing systems for identifying temporal relations between events heavily rely on hand-crafted features derived from event words and explicit temporal markers. Besides, less attention has been given to automatically learning con-textualized event representations or to finding complex interactions between events. In [9], we fill this gap in showing that a combination of rich event representations and interaction learning is essential to more accurate temporal relation classification. Specifically, we propose a method in which i) Recurrent Neural Networks (RNN) extract contextual information ii) character embeddings capture morpho-semantic features (e.g. tense, mood, aspect), and iii) a deep Convolutional Neural Network (CNN) finds out intricate interactions between events. We show that the proposed approach outperforms most existing systems on the commonly used dataset while using fully automatic feature extraction and simple local inference.

Phylogenetic Multi-Lingual Dependency Parsing Languages evolve and diverge over time. Their evolutionary history is often depicted in the shape of a phylogenetic tree. Assuming parsing models are representations of their languages grammars, their evolution should follow a structure similar to that of the phylo-genetic tree. In [7], drawing inspiration from multi-task learning, we make use of the phylogenetic tree to guide the learning of multilingual dependency parsers leverag-ing languages structural similarities. Experiments on data from the Universal Dependency project show that phylogenetic training is beneficial to low resourced languages and to well furnished languages families. As a side product of phylogenetic training, our model is able to perform zero-shot parsing of previously unseen languages.

7.2. Decentralized Learning

Trade-offs in Large-Scale Distributed Tuplewise Estimation and Learning The development of cluster computing frameworks has allowed practitioners to scale out various statistical estimation and machine learning algorithms with minimal programming effort. This is especially true for machine learning problems whose objective function is nicely separable across individual data points, such as classification and regression. In contrast, statistical learning tasks involving pairs (or more generally tuples) of data points-such as metric learning, clustering or ranking-do not lend themselves as easily to data-parallelism and in-memory computing. In [13], we investigate how to balance between statistical performance and computational efficiency in such distributed tuplewise statistical problems. We first propose a simple strategy based on occasionally repartitioning data across workers between parallel computation stages, where the number of repartition-ing steps rules the trade-off between accuracy and runtime. We then present some theoretical results highlighting the benefits brought by the proposed method in terms of variance reduction, and extend our results to design distributed stochastic gradient descent algorithms for tuplewise empirical risk minimization. Our results are supported by numerical experiments in pairwise statistical estimation and learning on synthetic and real-world datasets.

Who started this rumor? Quantifying the natural differential privacy guarantees of gossip protocols Gossip protocols, also called rumor spreading or epidemic protocols, are widely used to disseminate information in massive peer-to-peer networks. These protocols are often claimed to guarantee privacy because of the uncertainty they introduce on the node that started the dissemination. But is that claim really true? Can one indeed start a gossip and safely hide in the crowd? In [14], we study gossip protocols using a rigorous mathematical framework based on differential privacy to determine the extent to which the source of a gossip can be traceable. Considering the case of a complete graph in which a subset of the nodes are curious, we derive matching lower and upper bounds on differential privacy showing that some gossip protocols achieve strong privacy guarantees. Our results further reveal an interesting tension between privacy and dissemination speed: the standard “push” gossip protocol has very weak privacy guarantees, while the optimal guarantees are attained at the cost of a drastic increase in the spreading time. Yet, we show that it is possible to leverage the inherent randomness and partial observability of gossip protocols to achieve both fast dissemination speed and near-optimal privacy.

Fully Decentralized Joint Learning of Personalized Models and Collaboration Graphs In [15], we consider the fully decentralized machine learning scenario where many users with personal datasets collaborate to learn models through local peer-to-peer exchanges, without a central coordinator. We propose to train personalized models that leverage a collaboration graph describing the relationships between the users' personal tasks, which we learn jointly with the models. Our fully decentralized optimization procedure alternates between training nonlinear models given the graph in a greedy boosting manner, and updating the collaboration graph (with controlled sparsity) given the models. Throughout the process, users exchange messages only with a small number of peers (their direct neighbors in the graph and a few random users), ensuring that the procedure naturally scales to large numbers of users. We analyze the convergence rate, memory and communication complexity of our approach, and demonstrate its benefits compared to competing techniques on synthetic and real datasets.

Advances and Open Problems in Federated Learning Federated learning (FL) is a machine learning setting where many clients (e.g. mobile devices or whole organizations) collaboratively train a model under the orchestration of a central server (e.g. service provider), while keeping the training data decentralized. FL embodies the principles of focused data collection and minimization, and can mitigate many of the systemic privacy risks and costs resulting from traditional, centralized machine learning and data science approaches. Motivated by the explosive growth in FL research, we participated in a collaborative paper [18] that discusses recent advances and presents an extensive collection of open problems and challenges.

7.3. Privacy and Machine Learning

Private Protocols for U-Statistics in the Local Model and Beyond In [16], we study the problem of computing U-statistics of degree 2, i.e., quantities that come in the form of averages over pairs of data points, in the local model of differential privacy (LDP). The class of U-statistics covers many statistical estimates of interest, including Gini mean difference, Kendall's tau coefficient and Area under the ROC Curve (AUC), as well as empirical risk measures for machine learning problems such as ranking, clustering and metric learning. We first introduce an LDP protocol based on quantizing the data into bins and applying randomized response, which guarantees an ϵ-LDP estimate with a Mean Squared Error (MSE) of $O(1/\sqrt{n}\epsilon)$ under regularity assumptions on the U-statistic or the data distribution. We then propose a specialized protocol for AUC based on a novel use of hierarchical histograms that achieves MSE of $O(\alpha^3/n\epsilon^2)$ for arbitrary data distribution. We also show that 2-party secure computation allows to design a protocol with MSE of $O(1/n\epsilon^2)$, without any assumption on the kernel function or data distribution and with total communication linear in the number of users n. Finally, we evaluate the performance of our protocols through experiments on synthetic and real datasets.

Privacy-Preserving Adversarial Representation Learning in ASR: Reality or Illusion? In [11], we study Automatic Speech Recognition (ASR), a key technology in many services and applications. This typically requires user devices to send their speech data to the cloud for ASR decoding. As the speech signal carries a lot of information about the speaker, this raises serious privacy concerns. As a solution, an encoder may reside on each user device which performs local computations to anonymize the representation. In this paper, we focus on the protection of speaker identity and study the extent to which users can be recognized based on the encoded representation of their speech as obtained by a deep encoder-decoder architecture trained for ASR. Through speaker identification and verification experiments on the Librispeech corpus with open and closed sets of speakers, we show that the representations obtained from a standard architecture still carry a lot of information about speaker identity. We then propose to use adversarial training to learn representations that perform well in ASR while hiding speaker identity. Our results demonstrate that adversarial training dramatically reduces the closed-set classification accuracy, but this does not translate into increased open-set verification error hence into increased protection of the speaker identity in practice. We suggest several possible reasons behind this negative result.

Evaluating Voice Conversion-based Privacy Protection against Informed Attackers Speech signals are a rich source of speaker-related information including sensitive attributes like identity or accent. With a small amount of found speech data, such attributes can be extracted and modeled for malicious purposes like voice cloning, spoofing, etc. In [19], we investigate speaker anonymization strategies based on voice conversion. In contrast to prior evaluations, we argue that different types of attackers can be defined depending on the extent of their knowledge about the conversion scheme. We compare two frequency warping-based conversion methods and a deep learning based method in three attack scenarios. The utility of the converted speech is measured through the word error rate achieved by automatic speech recognition, while privacy protection is assessed by state-of-the-art speaker verification techniques (i-vectors and x-vectors). Our results show that voice conversion schemes are unable to effectively protect against an attacker that has extensive knowledge of the type of conversion and how it has been applied, but may provide some protection against less knowledgeable attackers.

7.4. Learning in Graphs

Correlation Clustering with Adaptive Similarity Queries In correlation clustering, we are given n objects together with a binary similarity score between each pair of them. The goal is to partition the objects into clusters so to minimise the disagreements with the scores. In [6], we investigate correlation clustering as an active learning problem: each similarity score can be learned by making a query, and the goal is to minimise both the disagreements and the total number of queries. On the one hand, we describe simple active learning algorithms, which provably achieve an almost optimal trade-off while giving cluster recovery guarantees, and we test them on different datasets. On the other hand, we prove information-theoretical bounds on the number of queries necessary to guarantee a prescribed disagreement bound. These results give a rich characterization of the trade-off between queries and clustering error.

Flattening a Hierarchical Clustering through Active Learning In [12], we investigate active learning by pairwise similarity over the leaves of trees originating from hierarchical clustering procedures. In the realizable setting, we provide a full characterization of the number of queries needed to achieve perfect reconstruction of the tree cut. In the non-realizable setting, we rely on known important-sampling procedures to obtain regret and query complexity bounds. Our algorithms come with theoretical guarantees on the statistical error and, more importantly, lend themselves to linear-time implementations in the relevant parameters of the problem. We discuss such implementations, prove running time guarantees for them, and present preliminary experiments on real-world datasets showing the compelling practical performance of our algorithms as compared to both passive learning and simple active learning baselines.

MaxHedge: Maximising a Maximum Online In [10], we introduce a new online learning framework where, at each trial, the learner is required to select a subset of actions from a given known action set. Each action is associated with an energy value, a reward and a cost. The sum of the energies of the actions selected cannot exceed a given energy budget. The goal is to maximise the cumulative profit, where the profit obtained on a single trial is defined as the difference between the maximum reward among the selected actions and the sum of their costs. Action energy values and the budget are known and fixed. All rewards and costs associated with each action change over time and are revealed at each trial only after the learner's selection of actions. Our framework encompasses several online learning problems where the environment changes over time; and the solution trades-off between minimising the costs and maximising the maximum reward of the selected subset of actions, while being constrained to an action energy budget. The algorithm that we propose is efficient, general and may be specialised to multiple natural online combinatorial problems.

Closed-loop cycles of experiment design, execution, and learning accelerate systems biology model development in yeast One of the most challenging tasks in modern science is the development of systems biology models: Existing models are often very complex but generally have low predictive performance. The construction of high-fidelity models will require hundreds/thousands of cycles of model improvement, yet few current systems biology research studies complete even a single cycle. In [2], we combined multiple software tools with integrated laboratory robotics to execute three cycles of model improvement of the prototypical eukaryotic cellular transformation, the yeast (Saccharomyces cerevisiae) diauxic shift. In the first cycle, a model outperforming the best previous diauxic shift model was developed using bioinformatic and systems biology tools. In the second cycle, the model was further improved using automatically planned experiments. In the third cycle, hypothesis-led experiments improved the model to a greater extent than achieved using high-throughput experiments. All of the experiments were formalized and communicated to a cloud laboratory automation system (Eve) for automatic execution, and the results stored on the semantic web for reuse. The final model adds a substantial amount of knowledge about the yeast diauxic shift: 92 genes (+45%), and 1 048 interactions (+147%). This knowledge is also relevant to understanding cancer, the immune system, and aging. We conclude that systems biology software tools can be combined and integrated with laboratory robots in closed-loop cycles.

7.5. Metric Learning

Metric learning is at the core of many algorithms for learning graphs. A new software has been published in the scikit-learn contrib repository (See the Software section).

Escaping the Curse of Dimensionality in Similarity Learning: Efficient Frank-Wolfe Algorithm and Generalization Bounds Similarity and metric learning provides a principled approach to construct a task-specific similarity from weakly supervised data. However, these methods are subject to the curse of dimensionality: as the number of features grows large, poor generalization is to be expected and training becomes intractable due to high computational and memory costs. In [3], we propose a similarity learning method that can efficiently deal with high-dimensional sparse data. This is achieved through a parameterization of similarity functions by convex combinations of sparse rank-one matrices, together with the use of a greedy approximate Frank-Wolfe algorithm which provides an efficient way to control the number of active features. We show that the convergence rate of the algorithm, as well as its time and memory complexity, are independent of the data dimension. We further provide a theoretical justification of our modeling choices through an analysis of the generalization error, which depends logarithmically on the sparsity of the solution rather than on the number of features. Our experiments on datasets with up to one million features demonstrate the ability of our approach to generalize well despite the high dimensionality as well as its superiority compared to several competing methods.

metric-learn: Metric Learning Algorithms in Python In [20], we present metric-learn, an open source Python package implementing supervised and weakly-supervised distance metric learning algorithms. As part of scikit-learn-contrib, it provides a unified interface compatible with scikit-learn which allows to easily perform cross-validation, model selection, and pipelining with other machine learning estimators. metric-learn is thoroughly tested and available on PyPi under the MIT licence.

7.6. Graph Algorithms

We collaborate with the Links project team on graph-based computations and evaluation in databases.

Dependency Weighted Aggregation on Factorized Databases In [17], we study a new class of aggregation problems, called dependency weighted aggregation. The underlying idea is to aggregate the answer tuples of a query while accounting for dependencies between them, where two tuples are considered dependent when they have the same value on some attribute. The main problem we are interested in is to compute the dependency weighted count of a conjunctive query. This aggregate can be seen as a form of weighted counting, where the weights of the answer tuples are computed by solving a linear program. This linear program enforces that dependent tuples are not over represented in the final weighted count. The dependency weighted count can be used to compute the s-measure, a measure that is used in data mining to estimate the frequency of a pattern in a graph database. Computing the dependency weighted count of a conjunctive query is NP-hard in general. In this paper, we show that this problem is actually tractable for a large class of structurally restricted conjunctive queries such as acyclic or bounded hypertree width queries. Our algorithm works on a factorized representation of the answer set, in order to avoid enumerating it exhaustively. Our technique produces a succinct representation of the weighting of the answers. It can be used to solve other dependency weighted aggregation tasks, such as computing the (dependency) weighted average of the value of an attribute in the answers set.

7.7. Learning and Speech Recognition

We have worked on privacy and machine learning for speech recognition (See Section 7.3). Additional results concern kernel method for speech recognition.

Kernel Approximation Methods for Speech Recognition In [4], we study the performance of kernel methods on the acoustic modeling task for automatic speech recognition, and compare their performance to deep neural networks (DNNs). To scale the kernel methods to large data sets, we use the random Fourier feature method of Rahimi and Recht (2007). We propose two novel techniques for improving the performance of kernel acoustic models. First, we propose a simple but effective feature selection method which reduces the number of random features required to attain a fixed level of performance. Second, we present a number of metrics which correlate strongly with speech recognition performance when computed on the heldout set; we attain improved performance by using these metrics to decide when to stop training. Additionally, we show that the linear bottleneck method of Sainath et al. (2013a) improves the performance of our kernel models significantly, in addition to speeding up training and making the models more compact. Leveraging these three methods, the kernel methods attain token error rates between 0.5% better and 0.1% worse than fully-connected DNNs across four speech recognition data sets, including the TIMIT and Broadcast News benchmark tasks.

MEPHYSTO Team

5. New Results

5.1. Traveling waves for some nonlocal 1D Gross-Pitaevskii equations with nonzero conditions at infinity

The nonlocal Gross-Pitaevskii equation is a model that appears naturally in several areas of quantum physics, for instance in the description of superfluids and in optics when dealing with thermo-optic materials because the thermal nonlinearity is usually highly nonlocal. A. de Laire and P. Mennuni have considered a nonlocal family of Gross–Pitaevskii equations in dimension one, and they have provided in [27] conditions on the nonlocal interaction such that there is a branch of traveling waves solutions with nonvanishing conditions at infinity. Moreover, they showed that the branch is orbitally stable. In this manner, this result generalizes known properties for the contact interaction given by a Dirac delta function. Their proof relies on the minimization of the energy at fixed momentum.

5.2. Numerical simulation of traveling waves for some nonlocal Gross-Pitaevskii equations with nonzero conditions at infinity in dimensions 1 and 2

As a follow-up of the previous result, P. Mennuni and G. Dujardin carried out numerical simulations of traveling waves for some nonlocal nonlinear Gross-Pitaevskii equations with nonzero conditions at infinity in dimensions 1 and 2. Using a numerical analogue of the minimization of the energy at fixed momentum, they used gradient methods with nonuniform fast Fourier transforms (to deal with the nonlocal terms numerically) to carry out significant numerical simulations to illustrate numerically the theoretical results and to discuss the hypotheses numerically. These results can be found in P. Mennuni's PhD manuscript [10].

5.3. The cubic Schrödinger regime of the Landau-Lifshitz equation with a strong easy-axis anisotropy

It is well-known that the dynamics of biaxial ferromagnets with a strong easy-axis anisotropy is essentially governed by the cubic Schrödinger equation. A. de Laire and P. Gravejat provided in [26] a rigorous justification to this observation. More precisely, they showed the convergence of the solutions to the Landau-Lifshitz equation for biaxial ferromagnets towards the solutions to the cubic Schrödinger equation in the regime of an easy-axis anisotropy. This result holds for solutions to the Landau–Lifshitz equation in high order Sobolev spaces. By introducing high order energy quantities with good symmetrization properties, they derived the convergence from the consistency of the Landau–Lifshitz equation with the Sine-Gordon equation by using well-tailored energy estimates.

In this regime, they additionally classified the one-dimensional solitons of the Landau-Lifshitz equation and quantified their convergence towards the solitons of the one-dimensional cubic Schrödinger equation.

5.4. The Cauchy problem for the Landau–Lifshitz–Gilbert equation in BMO and self-similar solutions

A. de Laire and S. Gutierrez established in [22] a global well-posedness result for the Landau–Lifshitz equation with Gilbert damping, provided that the BMO semi-norm of the initial data is small. As a consequence, they deduced the existence of self-similar solutions in any dimension. Moreover, in the one-dimensional case, they characterized the self-similar solutions when the initial data is given by some step function and established their stability. They also showed the existence of multiple solutions if the damping is strong enough.

5.5. Microscopic derivation of moving interfaces problems

In [15], M. Simon and her coauthors derive the porous medium equation from an interacting particle system which belongs to the family of kinetically constrained lattice gases. It was already proved in the literature that the macroscopic density profile is governed by the porous medium equation for initial densities uniformly bounded away from 0 and 1. Here we consider the more general case where the density can take those extreme values. The solutions display a richer behavior, like moving interfaces, finite speed of propagation and breaking of regularity. Since standard techniques cannot be straightforwardly applied, we present a way to generalize the relative entropy method, by involving approximations of solutions to the hydrodynamic equation, instead of exact solutions.

In [16], M. Simon and her coauthors study the hydrodynamic limit for a similar one-dimensional exclusion process but with an even more restricting dynamical constraint: this process with degenerate jump rates admits transient states, which it eventually leaves to reach an ergodic component if the initial macroscopic density is larger than a critical value, or one of its absorbing states otherwise. They show that, for initial profiles smooth enough and uniformly larger than the critical density, the macroscopic density profile evolves under the diffusive time scaling according to a fast diffusion equation. The first step in the proof is to show that the system typically reaches an ergodic component in subdiffusive time.

These two macroscopic behaviors belong to the class of moving interfaces problems, which are particularly hard to derive from the microscopic point of view.

5.6. Towards the weak KPE universality conjecture

In [32], P. Gonçalves, N. Perkowski and M. Simon derive the KPZ equation with boundary conditions, from an interacting particle system in contact with stochastic reservoirs, and they legitimate the choice done at the macroscopic level for the KPZ equation from the microscopic description of the system. This is more subtle than expected, because the boundary conditions do not behave canonically. The main challenge is to clarify the link between the macroscopic boundary effects and their atomic description.

5.7. Joule effect in chains of oscillators

In physics, the rotor chain has been investigated as an example of a system with two conserved quantities (angular momentum and energy), for which the thermal conductivity is finite (and therefore energy diffuses). Numerics shows an unexpected behaviour of the chain when the latter is connected at the boundaries to two thermostats, and a mechanical force imposes an average angular momentum at one boundary: the stationary temperature profile coincides with the values of the thermostats, but in the middle of the chain it raises to a much higher value. This behaviour is related to the presence of two conserved quantities and is sometimes referred to as Joule effect. Since the rotor model is too difficult to be treated analytically, T. Komorowski, S. Olla and M. Simon investigate in [25] the harmonic chain of oscillators, perturbed with a stochastic noise, which makes the heat transport diffusive, namely: the noise destroys the conservation law of the total momentum, but keeps the other two conservation laws (energy and stretch) intact. The boundaries of the chain are connected to two Langevin thermostats and an external force acting on one boundary puts the system in a non-equilibrium stationary state. The authors rigorously derive the Joule effect for a particular value of the noise intensity.

5.8. Quantum optics

In [18], S. De Bièvre and his co-authors introduce a new measure of the nonclassicality of the quantum states of an optical field, the so-called “ordering sensitivity” of the state, that measures the fluctuations of its Wigner function. This work is prolonged in two subsequent papers. In [24], S. De Bièvre and the same co-authors investigate a new class of quantum states they call the “Thermal Difference States” that can be generated by parametric down conversion. They investigate in particular the degree to which such states are nonclassical. In [34], S. De Bièvre and his postdoc A. Hertz, re-interpret the ordering sensitivity in terms of another physical property of the quantum states of an optical field, namely their quadrature coherence scale. It is shown in particular that a large such coherence scale is responsible for very fast environmental decoherence of the state.

5.9. Orbital stability

In [19], S. De Bièvre and S. Rota Nodari consider the orbital stability of relative equilibria of Hamiltonian dynamical systems on Banach spaces, in the presence of a multi-dimensional invariance group for the dynamics. They prove a persistence result for such relative equilibria, present a generalization of the Vakhitov-Kolokolov slope condition to this higher dimensional setting, and show how it allows to prove the local coercivity of the Lyapunov function, which in turn implies orbital stability. The method is applied to study the orbital stability of relative equilibria of nonlinear Schrödinger and Manakov equations. A comparison of their approach to the one by Grillakis-Shatah-Strauss is provided.

5.10. Exponential time-decay for discrete Fokker–Planck equations

In the research direction exposed in Section 3.3 , G. Dujardin and his coauthors proposed and studied in [21] several discrete versions of homogeneous and inhomogeneous one-dimensional Fokker-Planck equations. They proved in particular, for these discretizations of velocity and space, the exponential convergence to the equilibrium of the solutions, for time-continuous equations as well as for time-discrete equations. Their method uses new types of discrete Poincaré inequalities for a "two-direction" discretization of the derivative in velocity. For the inhomogeneous problem, they adapted for the very first time hypocoercive methods to the discrete level.

5.11. Energy preserving methods for nonlinear Schrödinger equations

G. Dujardin and his coauthors have revisited and extended relaxation methods for nonlinear Schrödinger equations (NLS). The classical relaxation method for NLS is an energy preserving method and a mass preserving method. Moreover, it is only linearly implicit. A first proof of the second order accuracy was achieved in [14]. Moreover, the method was extended to enable to treat noncubic nonlinearities, nonlocal nonlinearities, as well as rotation terms. The resulting methods are still energy preserving and mass preserving. Moreover, they are shown to have second order accuracy numerically. These new methods are compared with fully implicit, mass and energy preserving methods of Crank and Nicolson.

5.12. High order linearly implicit methods for evolution problems

In [31], I. Lacroix and G. Dujardin have developed a new class of numerical integration methods for evolution problems. This class contains methods of arbitrarily high order that only require the solution of a linear system per time step. For evolution ODEs (Cauchy problems), they give a constructive proof of existence for such arbitrarily high order methods. For evolution PDEs, they demonstrate numerically that these new methods can outperform high order methods from the literature on several test cases.

5.13. CLT for Circular beta-Ensembles at High Temperature

In [33], A. Hardy and G. Lambert have obtained a central limit theorem for the 2D Coulomb gas particle system constrained on a circle in the high temperature regime. An interesting feature is that the limiting variance interpolates between the Lebesgue L^2 norm, corresponding to the infinite temperature setting, and the Sobolev $H^{1/2}$ seminorm, corresponding to the zero temperature regime.

5.14. DLR equations and rigidity for the Sine-β process

The work [20] by A. Hardy and his collaborators, recently accepted for publication in Communications on Pure and Applied Mathematics, provides a "statistical physics" description of the sine-β process by means of Dobroshin-Lanford-Ruelle (DLR) equations. This basically allows to give a meaning to "the natural infinite configurations process on the real line in the 2D Coulomb interaction", provided there is a unique solution to the DLR equation which turns out to be true in this setting.

MODAL Project-Team

7. New Results

7.1. Axis 1: Data Units Selection in Statistics

Participant: Christophe Biernacki.

Usually, the data unit definition is fixed by the practitioner but it can happen that he/she hesitates between several data unit options. In this context, it is highlighted that it is possible to embed data unit selection into a classical model selection principle. The problem is introduced in a regression context before to focus on the model-based clustering and co-clustering context, for data of different kinds (continuous, count, categorical). This work was published in an international journal in 2018 and leads to a keynote as an invited speaker to the 12th Scientific Meeting Classification and Data Analysis Group Cassino (CLADAG 2019) in Italy [41].

It is a joint work with Alexandre Lourme from University of Bordeaux.

7.2. Axis 1: Model-Based Co-clustering for Ordinal Data of different dimensions

Participant: Christophe Biernacki.

This work has been motivated by a psychological survey on women affected by a breast tumor. Patients replied at different moments of their treatment to questionnaires with answers on ordinal scale. The questions relate to aspects of their life called dimensions. To assist the psychologists in analyzing the results, it is useful to emphasize a structure in the dataset. The clustering method achieves that by creating groups of individuals that are depicted by a representative of the group. From a psychological position, it is also useful to observe how questions may be grouped. This is why a clustering should also be performed on the features, which is called a co-clustering problem. However, gathering questions that are not related to the same dimension does not make sense from a psychologist stance. Therefore, the present work corresponds to perform a constrained co-clustering method aiming to prevent questions from different dimensions from getting assembled in a same column-cluster. In addition, evolution of co-clusters along time has been investigated. The method relies on a constrained Latent Block Model embedding a probability distribution for ordinal data. Parameter estimation relies on a Stochastic EM-algorithm associated to a Gibbs sampler, and the ICL-BIC criterion is used for selecting the numbers of co-clusters. The resulting work is now accepted in an international journal [28]. The related R package ordinalClust has been also written and has led to a specific preprint [73] now submitted to an international journal.

This is joint work with Margot Selosse (PhD student) and Julien Jacques, both from University of Lyon 2, and Florence Cousson-Gélie from University Paul Valéry Montpellier 3.

7.3. Axis 1: Model-based co-clustering for mixed type data

Participant: Christophe Biernacki.

Over decades, a lot of studies have shown the importance of clustering to emphasize groups of observations. More recently, due to the emergence of high-dimensional datasets with a huge number of features, co-clustering techniques have emerged and proposed several methods for simultaneously producing groups of observations and features. By synthesizing the dataset in blocks (the crossing of a row-cluster and a column-cluster), this technique can sometimes summarize better the data and its inherent structure. The Latent Block Model (LBM) is a well-known method for performing a co-clustering. However, recently, contexts with features of different types (here called mixed type datasets) are becoming more common. Unfortunately, the LBM is not directly applicable on this kind of dataset. The present work extends the usual LBM to the so-called Multiple Latent Block Model (MLBM) which is able to handle mixed type datasets. The inference is done through a Stochastic EM-algorithm embedding a Gibbs sampler and model selection criterion is defined to choose the number of row and column clusters. This method was successfully used on simulated and real datasets. This work is now accepted in an international journal [29].

This is joint work with Margot Selosse (PhD student) and Julien Jacques, both from University of Lyon 2.

7.4. Axis 1: Relaxing the Identically Distributed Assumption in Gaussian Co-Clustering for High Dimensional Data

Participant: Christophe Biernacki.

A co-clustering model for continuous data that relaxes the identically distributed assumption within blocks of traditional co-clustering is presented. The proposed model, although allowing more flexibility, still maintains the very high degree of parsimony achieved by traditional co-clustering. A stochastic EM algorithm along with a Gibbs sampler is used for parameter estimation and an ICL criterion is used for model selection. Simulated and real datasets are used for illustration and comparison with traditional co-clustering. This work has been submitted to an international journal [63].

This is a joint work with Michael Gallaugher (PhD student) and Paul McNicholas, both from McMaster University (Canada). Michael Gallaugher visited Modal for three months in 2018.

7.5. Axis 1: Gaussian-based visualization of Gaussian and non-Gaussian model-based clustering

Participants: Christophe Biernacki, Vincent Vandewalle.

A generic method is introduced to visualize in a Gaussian-like way, and onto R^2, results of Gaussian or non-Gaussian model-based clustering. The key point is to explicitly force a spherical Gaussian mixture visualization to inherit from the within cluster overlap which is present in the initial clustering mixture. The result is a particularly user-friendly draw of the clusters, allowing any practitioner to have a thorough overview of the potentially complex clustering result. An entropic measure allows us to inform of the quality of the drawn overlap, in comparison to the true one in the initial space. The proposed method is illustrated on four real data sets of different types (categorical, mixed, functional and network) and is implemented on the R package ClusVis. This work is now in minor revision for an international journal [54]. It has also led to an invited talk to an international conference [42], and several other invitations (the workshop "Advances in data science for big and complex data" at Université Paris-Dauphine in January and the seminary of the Probability and Statistics team of the University Nice Sophia-Antipolis in November).

This is a joint work with Matthieu Marbac from ENSAI.

7.6. Axis 1: Co-clustering: A versatile way to perform clustering

Participant: Christophe Biernacki.

Standard model-based clustering is known to be very efficient for low dimensional data sets, but it fails for properly addressing high dimension (HD) ones, where it suffers from both statistical and computational drawbacks. In order to counterbalance this curse of dimensionality, some proposals have been made to take into account redundancy and features utility, but related models are not suitable for too many variables. We advocate that the latent bloc model, a probabilistic model for co-clustering, is of particular interest to perform HD clustering of individuals even if it is not its primary function. We illustrate in an empirical manner the trade-off bias-variance of the co-clustering strategy in scenarii involving HD fundaments (correlated variables, irrelevant variables) and show the ability of co-clustering to outperform simple mixture row-clustering. An early version of this work has been presented to an national conference with international audience [46].

We also co-organized a special session to an international conference [45] to discuss the potential links between deterministic methods for co-clustering (based on a metric and computer science procedure) or probabilistic methods for co-clustering (mainly based on mixture models). It was the opportunity to gather related communities which are often distinct.

All are joint works with Christine Keribin from Université Paris-Sud.

7.7. Axis 1: Dealing with missing data in model-based clustering through a MNAR model

Participants: Christophe Biernacki, Fabien Laporte.

Since the 90s, model-based clustering is largely used to classify data. Nowadays, with the increase of available data, missing values are more frequent. Traditional ways to deal with them consist in obtaining a filled data set, either by discarding missing values or by imputing them. In the first case, some information is lost; in the second case, the final clustering purpose is not taken into account through the imputation step. Thus, both solutions risk to blur the clustering estimation result. Alternatively, we defend the need to embed the missingness mechanism directly within the clustering modeling step. There exists three types of missing data: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). In all situations logistic regression is proposed as a natural and flexible candidate model. In particular, its flexibility property allows us to design some meaningful parsimonious variants, as dependency on missing values or dependency on the cluster label. In this unified context, standard model selection criteria can be used to select between such different missing data mechanisms, simultaneously with the number of clusters. Practical interest of our proposal is illustrated on data derived from medical studies suffering from many missing data. This work has been presented as an invited speaker to an international conference [31]. It has also been presented at a national conference with international audience [47] and as a poster to the international Working Group on Model-Based Clustering [69]. Currently, a preprint is being finalized for submission to an international journal.

It is a joint work with Gilles Celeux from Inria Saclay and Julie Josse from Ecole Polytechnique.

7.8. Axis 1: Organized Co-Clustering for textual data synthesis

Participant: Christophe Biernacki.

Recently, different studies have demonstrated the interest of co-clustering, which simultaneously produces clusters of lines and columns. The present work introduces a novel co-clustering model for parsimoniously summarizing textual data in documents × terms format. Besides highlighting homogeneous coclusters - as other existing algorithms do - we also distinguish noisy coclusters from significant ones, which is particularly useful for sparse documents × term matrices. Furthermore, our model proposes a structure among the significant coclusters and thus obtains a better interpretability to the user. By forcing a structure through row-clusters and column-clusters, this approach is competitive in terms of documents clustering, and offers user-friendly results. The algorithm derived for the proposed method is a Stochastic EM algorithm embedding a Gibbs sampling step and the Poisson distribution. A paper is currently in revision in an international journal [72].

This is joint work with Margot Selosse (PhD student) and Julien Jacques, both from University of Lyon 2.

7.9. Axis 1: Model-Based Co-clustering with Co-variables

Participant: Serge Iovleff.

This work has been motivated by an epidemiological and genetic survey of malaria disease in Senegal. Data were collected between 1990 and 2008. It is based on a latent block model taking into account the problem of grouping variables and clustering individuals by integrating information given by a set of co-variables. Numerical experiments on simulated data sets and an application on real genetic data highlight the interest of this approach. An article has been submitted to *Journal of Classification* and should incorporate “Major Revisions”.

7.10. Axis 1: Linking canonical and spectral clustering

Participants: Christophe Biernacki, Vincent Vandewalle.

It is a recent work aiming at defining a mathematical bridge between classical model-based clustering and classical spectral clustering. Interest of such a prospect is to be able to compare both methods through the rigorous scheme of model selection paradigm. It is still an ongoing work, with several short working papers.

It is a joint work with Alexandre Lourme from University of Bordeaux.

7.11. Axis 1: Predictive clustering

Participants: Christophe Biernacki, Vincent Vandewalle.

Many data, for instance in biostatistics, contain some sets of variables which permit evaluating unobserved traits of the subjects (e.g., we ask question about how many pizzas, hamburgers, chips... are eaten to know how healthy are the food habits of the subjects). Moreover, we often want to measure the relations between these unobserved traits and some target variables (e.g., obesity). Thus, a two-steps procedure is often used: first, a clustering of the observations is performed on the sets of variables related to the same topic; second, the predictive model is fitted by plugging the estimated partitions as covariates. Generally, the estimated partitions are not exactly equal to the true ones. We investigate the impact of these measurement errors on the estimators of the regression parameters, and we explain when this two-steps procedure is consistent. We also present a specific EM algorithm which simultaneously estimates the parameters of the clustering and predictive models. It is an ongoing work.

It is a joint work with Matthieu Marbac from ENSAI and Mohammed Sedki from University Paris-Sud.

7.12. Axis 1: Ranking and synchronization from pairwise measurements via SVD

Participant: Hemant Tyagi.

Given a measurement graph $G = ([n], E)$ and an unknown signal $r \in R^n$, we investigate algorithms for recovering r from pairwise measurements of the form $r_i - r_j$; $\{i, j\} \in E$. This problem arises in a variety of applications, such as ranking teams in sports data and time synchronization of distributed networks. Framed in the context of ranking, the task is to recover the ranking of n teams (induced by r) given a small subset of noisy pairwise rank offsets. We propose a simple SVD-based algorithmic pipeline for both the problem of time synchronization and ranking. We provide a detailed theoretical analysis in terms of robustness against both sampling sparsity and noise perturbations with outliers, using results from matrix perturbation and random matrix theory. Our theoretical findings are complemented by a detailed set of numerical experiments on both synthetic and real data, showcasing the competitiveness of our proposed algorithms with other state-of-the-art methods.

This is joint work with Alexandre d'Aspremont (CNRS & ENS, Paris) and Mihai Cucuringu (University of Oxford, UK) and is available as a preprint [61].

7.13. Axis 1: SPONGE: A generalized eigenproblem for clustering signed networks

Participant: Hemant Tyagi.

We introduce a principled and theoretically sound spectral method for k-way clustering in signed graphs, where the affinity measure between nodes takes either positive or negative values. Our approach is motivated by social balance theory, where the task of clustering aims to decompose the network into disjoint groups, such that individuals within the same group are connected by as many positive edges as possible, while individuals from different groups are connected by as many negative edges as possible. Our algorithm relies on a generalized eigenproblem formulation inspired by recent work on constrained clustering. We provide theoretical guarantees for our approach in the setting of a signed stochastic block model, by leveraging tools from matrix perturbation theory and random matrix theory. An extensive set of numerical experiments on both synthetic and real data shows that our approach compares favorably with state-of-the-art methods for signed clustering, especially for large number of clusters and sparse measurement graphs.

This is joint work with Mihai Cucuringu (University of Oxford, UK), Peter Davies (University of Warwick, UK) and Aldo Glielmo (Imperial College, London, UK) and was mostly done while Hemant Tyagi was affiliated to the Alan Turing Institute. It was published in the proceedings of an international conference [32].

7.14. Axis 2: Multi-kernel unmixing and super-resolution using the Modified Matrix Pencil method

Participant: Hemant Tyagi.

Consider L groups of point sources or spike trains, with the l^{th} group represented by $x_l(t)$. For a function $g : R \to R$, let $g_l(t) = g(t/\mu_l)$ denote a point spread function with scale $\mu_l > 0$, and with $\mu_1 < \cdots < \mu_L$. With $y(t) = \sum_{l=1}^{L} (g_l \bar{\star} x_l)(t)$, our goal is to recover the source parameters given samples of y, or given the Fourier samples of y. This problem is a generalization of the usual super-resolution setup wherein $L = 1$; we call this the multi-kernel unmixing super-resolution problem. Assuming access to Fourier samples of y, we derive an algorithm for this problem for estimating the source parameters of each group, along with precise non-asymptotic guarantees. Our approach involves estimating the group parameters sequentially in the order of increasing scale parameters, i.e., from group 1 to L. In particular, the estimation process at stage $1 \leq l \leq L$ involves (i) carefully sampling the tail of the Fourier transform of y, (ii) a *deflation* step wherein we subtract the contribution of the groups processed thus far from the obtained Fourier samples, and (iii) applying Moitra's modified Matrix Pencil method on a deconvolved version of the samples in (ii).

This is joint work with Stephane Chretien (National Physical Laboratory, UK & Alan Turing Institute, London) and was mostly done while Hemant Tyagi was affiliated to the Alan Turing Institute. It is currently under revision in an international journal and is available as a preprint [56].

7.15. Axis 2: Provably robust estimation of modulo 1 samples of a smooth function with applications to phase unwrapping

Participant: Hemant Tyagi.

Consider an unknown smooth function $f : [0,1]^d \to R$, and assume we are given n noisy mod 1 samples of f, i.e., $y_i = (f(x_i) + \eta_i) \mod 1$, for $x_i \in [0,1]^d$, where η_i denotes the noise. Given the samples $(x_i, y_i)_{i=1}^n$, our goal is to recover smooth, robust estimates of the clean samples $f(x_i) \mod 1$. We formulate a natural approach for solving this problem, which works with angular embeddings of the noisy mod 1 samples over the unit circle, inspired by the angular synchronization framework. This amounts to solving a smoothness regularized least-squares problem – a quadratically constrained quadratic program (QCQP) – where the variables are constrained to lie on the unit circle. Our proposed approach is based on solving its relaxation, which is a *trust-region sub-problem* and hence solvable efficiently. We provide theoretical guarantees demonstrating its robustness to noise for adversarial, as well as random Gaussian and Bernoulli noise models. To the best of our knowledge, these are the first such theoretical results for this problem. We demonstrate the robustness and efficiency of our proposed approach via extensive numerical simulations on synthetic data, along with a simple least-squares based solution for the unwrapping stage, that recovers the original samples of f (up to a global shift). It is shown to perform well at high levels of noise, when taking as input the denoised modulo 1 samples. Finally, we also consider two other approaches for denoising the modulo 1 samples that leverage tools from Riemannian optimization on manifolds, including a Burer-Monteiro approach for a semidefinite programming relaxation of our formulation. For the two-dimensional version of the problem, which has applications in synthetic aperture radar interferometry (InSAR), we are able to solve instances of real-world data with a million sample points in under 10 seconds, on a personal laptop.

This is joint work with Mihai Cucuringu (University of Oxford, UK) and was mostly done while Hemant Tyagi was affiliated to the Alan Turing Institute. It has been accepted to appear (after minor revision) in an international journal, and is available as a preprint [60].

7.16. Axis 2: Learning general sparse additive models from point queries in high dimensions

Participant: Hemant Tyagi.

We consider the problem of learning a d-variate function f defined on the cube $[-1,1]^d \subset R^d$, where the algorithm is assumed to have black box access to samples of f within this domain. Denote $S_r; r = 1, \cdots, r_0$ to be sets consisting of unknown r-wise interactions amongst the coordinate variables. We then focus on the setting where f has an additive structure, i.e., it can be represented as

$$f = \sum_{j \in \S_1} \phi_j + \sum_{j \in S_2} \phi_j + \cdots + \sum_{j \in S_{r_0}} \phi_j,$$

where each ϕ_j; $j \in S_r$ is at most r-variate for $1 \leq r \leq r_0$. We derive randomized algorithms that query f at carefully constructed set of points, and exactly recover each S_r with high probability. In contrary to the previous work, our analysis does not rely on numerical approximation of derivatives by finite order differences.

This is joint work with Jan Vybiral (Czech Technical University, Prague) and was mostly done while Hemant Tyagi was affiliated to the Alan Turing Institute. It has now been published in an international journal [30].

7.17. Axis 2: Sparse non-negative super-resolution - simplified and stabilized

Participant: Hemant Tyagi.

The convolution of a discrete measure, $x = \sum_{i=1}^{k} a_i \delta_{t_i}$, with a local window function, $\phi(s-t)$, is a common model for a measurement device whose resolution is substantially lower than that of the objects being observed. Super-resolution concerns localising the point sources with an accuracy beyond the essential support of $\phi(s-t)$, typically from m noisy samples of the convolution output. We consider the setting of x being non-negative and seek to characterise all non-negative measures approximately consistent with the samples. We first show that x is the unique non-negative measure consistent with the samples provided the samples are exact, and $m \geq 2k+1$ samples are available, and $\phi(s-t)$ generates a Chebyshev system. This is independent of how close the sample locations are and *does not rely on any regulariser beyond non-negativity*; as such, it extends and clarifies the work by Schiebinger et al. and De Castro et al., who achieve the same results but require a total variation regulariser, which we show is unnecessary. Moreover, we establish stability results in the setting where the samples are corrupted with noise. The main innovation of these results is that non-negativity alone is sufficient to localise point sources beyond the essential sensor resolution.

This is joint work with Armin Eftekhari (EPFL, Switzerland), Jared Tanner (University of Oxford, UK), Andrew Thompson (National Physical Laboratory, UK), Bogdan Toader (University of Oxford, UK) and was mostly done while Hemant Tyagi was affiliated to the Alan Turing Institute. It has now been published in an international journal [24].

7.18. Axis 2: Pseudo-Bayesian learning with kernel Fourier transform as prior

Participant: Pascal Germain.

We revisit the kernel random Fourier features (RFF) method through the lens of the PAC-Bayesian theory. While the primary goal of RFF is to approximate a kernel, we look at the Fourier transform as a prior distribution over trigonometric hypotheses. It naturally suggests learning a posterior on these hypotheses. We derive generalization bounds that are optimized by learning a pseudo-posterior obtained from a closed-form expression, and corresponding learning algorithms.

This joint work with Emilie Morvant from Université Jean Monnet de Saint-Etienne (France), and Gaël Letarte from Université Laval (Québec, Canada) has been initiated in 2018 when Gaël Letarte was doing an internship at Inria, and led to a publication in the proceedings of AISTATS 2019 conference [36]. The same work has been prensented as a poster in the "Workshop on Machine Learning with guarantees @ NeurIPS 2019".

An extension of this work, co-authored with Léo Gautheron, Amaury Habrard, Marc Sebban, and Valentina Zantedeschi – all from Université Jean Monnet de Saint-Etienne – has been presented at the national conference CAp 2019 [44]. It is also the topic of a technical report [64].

7.19. Axis 2: PAC-Bayesian binary activated deep neural networks

Participant: Pascal Germain, Benjamin Guedj

We present a comprehensive study of multilayer neural networks with binary activation, relying on the PAC-Bayesian theory. Our contributions are twofold: (i) we develop an end-to-end framework to train a binary activated deep neural network, overcoming the fact that binary activation function is non-differentiable; (ii) we provide nonvacuous PAC-Bayesian generalization bounds for binary activated deep neural networks. Noteworthy, our results are obtained by minimizing the expected loss of an architecture-dependent aggregation of binary activated deep neural networks. The performance of our approach is assessed on a thorough numerical experiment protocol on real-life datasets. This work has been published in the proceedings of NeurIPS 2019 conference [35].

It is a joint work with Gaël Letarte and François Laviolette, from Université Laval (Québec, Canada).

7.20. Axis 2: Improved PAC-Bayesian Bounds for Linear Regression

Participant: Pascal Germain, Vera Shalaeva

We improve the PAC-Bayesian error bound for linear regression provided in the literature. The improvements are two-fold. First, the proposed error bound is tighter, and converges to the generalization loss with a well-chosen temperature parameter. Second, the error bound also holds for training data that are not independently sampled. In particular, the error bound applies to certain time series generated by well-known classes of dynamical models, such as ARX models.

It is a joint work with Mihaly Petreczky and Alireza Fakhrizadeh Esfahani from Université de Lille. It has been accepted for publication as part of the AAAI 2020 conference [38].

7.21. Axis 2: Multiview Boosting by controlling the diversity and the accuracy of view-specific voters

Participant: Pascal Germain

We present a comprehensive study of multilayer neural networks with binary activation, relying on the PAC-Bayesian We propose a boosting based multiview learning algorithm which iteratively learns i) weights over view-specific voters capturing view-specific information; and ii) weights over views by optimizing a PAC-Bayes multiview C-Bound that takes into account the accuracy of view-specific classifiers and the diversity between the views. We derive a generalization bound for this strategy following the PAC-Bayes theory which is a suitable tool to deal with models expressed as weighted combination over a set of voters.

It is a joint work with Emilie Morvant from Université Jean Monnet de Saint-Etienne and with Massih-Reza Amini of Université de Grenoble, and with Anil Goyal affiliated to both institutions. This work has been published in the journal Neurocomputing [26].

7.22. Axis 2: PAC-Bayes and Domain Adaptation

Participant: Pascal Germain

In machine learning, Domain Adaptation (DA) arises when the distribution generating the test (target) data differs from the one generating the learning (source) data. It is well known that DA is a hard task even under strong assumptions, among which the covariate-shift where the source and target distributions diverge only in their marginals, i.e. they have the same labeling function. Another popular approach is to consider a hypothesis class that moves closer the two distributions while implying a low-error for both tasks. This is a VC-dim approach that restricts the complexity of a hypothesis class in order to get good generalization. Instead, we propose a PAC-Bayesian approach that seeks for suitable weights to be given to each hypothesis in order to build a majority vote. We prove a new DA bound in the PAC-Bayesian context. This leads us to design the first DA-PAC-Bayesian algorithm based on the minimization of the proposed bound. Doing so, we seek for a ρ-weighted majority vote that takes into account a trade-off between three quantities. The first two quantities being, as usual in the PAC-Bayesian approach, (a) the complexity of the majority vote (measured by a Kullback-Leibler divergence) and (b) its empirical risk (measured by the ρ-average errors on the source sample). The third quantity is (c) the capacity of the majority vote to distinguish some structural difference between the source and target samples.

This work has been published in the journal Neurocomputing [25].

It is a joint work with Emilie Morvant and Amaury Habrard from Université Jean Monnet de Saint-Etienne (France), and with François Laviolette from Université Laval (Québec, Canada).

7.23. Axis 2: Interpreting Neural Networks as Majority Votes through the PAC-Bayesian Theory

Participant: Pascal Germain, Paul Viallard

We propose a PAC-Bayesian theoretical study of the two-phase learning procedure of a neural network introduced by Kawaguchi et al. (2017). In this procedure, a network is expressed as a weighted combination of all the paths of the network (from the input layer to the output one), that we reformulate as a PAC-Bayesian majority vote. Starting from this observation, their learning procedure consists in (1) learning “prior” network for fixing some parameters, then (2) learning a “posterior” network by only allowing a modification of the weights over the paths of the prior network. This allows us to derive a PAC-Bayesian generalization bound that involves the empirical individual risks of the paths (known as the Gibbs risk) and the empirical diversity between pairs of paths. Note that similarly to classical PAC-Bayesian bounds, our result involves a KL-divergence term between a “prior” network and the "posterior" network. We show that this term is computable by dynamic programming without assuming any distribution on the network weights.

This early result has been accepted as a poster presentation in the international workshop “Workshop on Machine Learning with guarantees @ NeurIPS 2019” [50].

This is a joint work with researchers from Université Jean Monnet de Saint-Etienne: Amaury Habrard, Emilie Morvant, and Rémi Emonet.

7.24. Axis 2: Still no free lunches: the price to pay for tighter PAC-Bayes bounds

Participant: Benjamin Guedj

“No free lunch” results state the impossibility of obtaining meaningful bounds on the error of a learning algorithm without prior assumptions and modelling. Some models are expensive (strong assumptions, such as as subgaussian tails), others are cheap (simply finite variance). As it is well known, the more you pay, the more you get: in other words, the most expensive models yield the more interesting bounds. Recent advances in robust statistics have investigated procedures to obtain tight bounds while keeping the cost minimal. The present paper explores and exhibits what the limits are for obtaining tight PAC-Bayes bounds in a robust setting for cheap models, addressing the question: is PAC-Bayes good value for money?

Joint work with Louis Pujol (Université Paris-Saclay). Available as a preprint: [68]

7.25. Axis 2: PAC-Bayesian Contrastive Unsupervised Representation Learning

Participant: Benjamin Guedj, Pascal Germain

Contrastive unsupervised representation learning (CURL) is the state-of-the-art technique to learn representations (as a set of features) from unlabelled data. While CURL has collected several empirical successes recently, theoretical understanding of its performance was still missing. In a recent work, Arora et al. (2019) provide the first generalisation bounds for CURL, relying on a Rademacher complexity. We extend their framework to the flexible PAC-Bayes setting, allowing to deal with the non-iid setting. We present PAC-Bayesian generalisation bounds for CURL, which are then used to derive a new representation learning algorithm. Numerical experiments on real-life datasets illustrate that our algorithm achieves competitive accuracy, and yields generalisation bounds with non-vacuous values.

Joint work with Kento Nozawa (University of Tokyo & RIKEN). Available as a preprint: [71]

7.26. Axis 2: Sequential Learning of Principal Curves: Summarizing Data Streams on the Fly

Participant: Benjamin Guedj

When confronted with massive data streams, summarizing data with dimension reduction methods such as PCA raises theoretical and algorithmic pitfalls. Principal curves act as a nonlinear generalization of PCA and the present paper proposes a novel algorithm to automatically and sequentially learn principal curves from data streams. We show that our procedure is supported by regret bounds with optimal sublinear remainder terms. A greedy local search implementation (called `slpc`, for Sequential Learning Principal Curves) that incorporates both sleeping experts and multi-armed bandit ingredients is presented, along with its regret computation and performance on synthetic and real-life data.

Joint work with Le Li (Université d'Angers & iAdvize). Available as a preprint: [67]

7.27. Axis 2: PAC-Bayes Un-Expected Bernstein Inequality

Participant: Benjamin Guedj

We present a new PAC-Bayesian generalization bound. Standard bounds contain a $\sqrt{L_n \cdot KL/n}$ complexity term which dominates unless L_n, the empirical error of the learning algorithm's randomized predictions, vanishes. We manage to replace L_n by a term which vanishes in many more situations, essentially whenever the employed learning algorithm is sufficiently stable on the dataset at hand. Our new bound consistently beats state-of-the-art bounds both on a toy example and on UCI datasets (with large enough n). Theoretically, unlike existing bounds, our new bound can be expected to converge to 0 faster whenever a Bernstein/Tsybakov condition holds, thus connecting PAC-Bayesian generalization and *excess risk* bounds—for the latter it has long been known that faster convergence can be obtained under Bernstein conditions. Our main technical tool is a new concentration inequality which is like Bernstein's but with X^2 taken outside its expectation.

Joint work with Peter Grünwald (CWI), Zakaria Mhammedi (Australian National University).

This work has been accepted at NeurIPS 2019, will be presented as a poster in the main conference and as a oral in the workshop "Machine Learning with guarantees", and is included in the proceedings of NeurIPS 2019.

Published: [37]

7.28. Axis 2: Attributing and Referencing (Research) Software: Best Practices and Outlook from Inria

Participant: Benjamin Guedj

Software is a fundamental pillar of modern scientiic research, not only in computer science, but actually across all elds and disciplines. However, there is a lack of adequate means to cite and reference software, for many reasons. An obvious rst reason is software authorship, which can range from a single developer to a whole team, and can even vary in time. The panorama is even more complex than that, because many roles can be involved in software development: software architect, coder, debugger, tester, team manager, and so on. Arguably, the researchers who have invented the key algorithms underlying the software can also claim a part of the authorship. And there are many other reasons that make this issue complex. We provide in this paper a contribution to the ongoing eeorts to develop proper guidelines and recommendations for software citation, building upon the internal experience of Inria, the French research institute for digital sciences. As a central contribution, we make three key recommendations. (1) We propose a richer taxonomy for software contributions with a qualitative scale. (2) We claim that it is essential to put the human at the heart of the evaluation. And (3) we propose to distinguish citation from reference.

Joint work with Pierre Alliez, Roberto Di Cosmo, Alain Girault, Mohand-Said Hacid, Arnaud Legrand, Nicolas Rougier (Inria).

This work has been published in the journal Computing in Science and Engineering.

Published: [14]

7.29. Axis 2: Revisiting clustering as matrix factorisation on the Stiefel manifold

Participant: Benjamin Guedj

This paper studies clustering for possibly high dimensional data (*e.g.* images, time series, gene expression data, and many other settings), and rephrase it as low rank matrix estimation in the PAC-Bayesian framework. Our approach leverages the well known Burer-Monteiro factorisation strategy from large scale optimisation, in the context of low rank estimation. Moreover, our Burer-Monteiro factors are shown to lie on a Stiefel manifold. We propose a new generalized Bayesian estimator for this problem and prove novel prediction bounds for clustering. We also devise a componentwise Langevin sampler on the Stiefel manifold to compute this estimator.

Joint work with Stéphane Chrétien (Université Lyon-2). Available as a preprint: [55]

7.30. Axis 2: A Primer on PAC-Bayesian Learning

Participant: Benjamin Guedj

This survey on PAC-Bayesian learning has been the backbone to a successful proposal for an ICML 2019 plenary tutorial.

Generalised Bayesian learning algorithms are increasingly popular in machine learning, due to their PAC generalisation properties and flexibility. The present paper aims at providing a self-contained survey on the resulting PAC-Bayes framework and some of its main theoretical and algorithmic developments.

This work has been published in the proceedings of the French Mathematical Society. Published as [66].

7.31. Axis 2: Perturbed Model Validation: A New Framework to Validate Model Relevance

Participant: Benjamin Guedj

This paper introduces Perturbed Model Validation (PMV), a new technique to validate model relevance and detect overfitting or underfitting. PMV operates by injecting noise to the training data, re-training the model against the perturbed data, then using the training accuracy decrease rate to assess model relevance. A larger decrease rate indicates better concept-hypothesis fit. We realise PMV by perturbing labels to inject noise, and evaluate PMV on four real-world datasets (breast cancer, adult, connect-4, and MNIST) and nine synthetic datasets in the classification setting. The results reveal that PMV selects models more precisely and in a more stable way than cross-validation, and effectively detects both overfitting and underfitting.

It is a joint work with Jie Zhang, Earl Barr, John Shawe-Taylor (all with UCL), and Mark Harman (UCL & Facebook). Available as a preprint: [75].

7.32. Axis 2: Decentralized learning with budgeted network load using Gaussian copulas and classifier ensembles

Participant: Benjamin Guedj

We examine a network of learners which address the same classification task but must learn from different data sets. The learners cannot share data but instead share their models. Models are shared only one time so as to preserve the network load. We introduce DELCO (standing for Decentralized Ensemble Learning with COpulas), a new approach allowing to aggregate the predictions of the classifiers trained by each learner. The proposed method aggregates the base classifiers using a probabilistic model relying on Gaussian copulas. Experiments on logistic regressor ensembles demonstrate competing accuracy and increased robustness in case of dependent classifiers. A companion python implementation is available online.

Joint work with John Klein, Olivier Colot, Mahmoud Albardan (Université de Lille).

This work has been published in the proceedings of ECML-PKDD 2019, as part (oral presentation) of the workshop Decentralized Machine Learning at the Edge. Published: [34]

7.33. Axis 2: Online k-means Clustering

Participant: Benjamin Guedj

We study the problem of online clustering where a clustering algorithm has to assign a new point that arrives to one of k clusters. The specific formulation we use is the k-means objective: At each time step the algorithm has to maintain a set of k candidate centers and the loss incurred is the squared distance between the new point and the closest center. The goal is to minimize regret with respect to the best solution to the k-means objective ($\mathcal{C}$) in hindsight. We show that provided the data lies in a bounded region, an implementation of the Multiplicative Weights Update Algorithm (*MWUA*) using a discretized grid achieves a regret bound of $\widetilde{O}(\sqrt{T})$ in expectation. We also present an online-to-offline reduction that shows that an efficient no-regret online algorithm (despite being allowed to choose a different set of candidate centres at each round) implies an offline efficient algorithm for the k-means problem. In light of this hardness, we consider the slightly weaker requirement of comparing regret with respect to $(1+\epsilon)\mathcal{C}$ and present a no-regret algorithm with runtime $O\left(T(\mathrm{poly}(\log(T), k, d, 1/\epsilon)^{k(d+O(1))}\right)$. Our algorithm is based on maintaining an incremental coreset and an adaptive variant of the *MWUA*. We show that naïve online algorithms, such as *Follow The Leader*, fail to produce sublinear regret in the worst case. We also report preliminary experiments with synthetic and real-world data.

Joint work with Varun Kanade, Guy Rom (University of Oxford), Vincent Cohen-Addad (CNRS). Available as a preprint: [57]

7.34. Axis 2: Non-linear aggregation of filters to improve image denoising

Participant: Benjamin Guedj

We introduce a novel aggregation method to efficiently perform image denoising. Preliminary filters are aggregated in a non-linear fashion, using a new metric of pixel proximity based on how the pool of filters reaches a consensus. We provide a theoretical bound to support our aggregation scheme, its numerical performance is illustrated and we show that the aggregate significantly outperforms each of the preliminary filters.

Joint work with Juliette Rengot (Ecole des Ponts).

This work has been accepted at the Computing Conference 2020 (July 2020, London, UK) and will be included in the proceedings. Published: [33]

7.35. Axis 2: Multiple change-points detection with reproducing kernels

Participant: Alain Celisse

We tackle the change-point problem with data belonging to a general set. We build a penalty for choosing the number of change-points in the kernel-based method of Harchaoui and Cappé (2007). This penalty generalizes the one proposed by Lebarbier (2005) for a one-dimensional signal changing only through its mean. We prove a non-asymptotic oracle inequality for the proposed method, thanks to a new concentration result for some function of Hilbert-space valued random variables. Experiments on synthetic and real data illustrate the accuracy of our method, showing that it can detect changes in the whole distribution of data, even when the mean and variance are constant.

Joint work with Sylvain Arlot (Orsay) and Zaïd Harchaoui (Seattle). This work has been accepted in JMLR [15].

7.36. Axis 2: Analysis of early stopping rules based on discrepancy principle

Participant: Alain Celisse

We describe a general unified framework for analyzing the statistical performance of early stopping rules based on the minimum discrepancy principle (DP). Finite-sample bounds such as deviation or oracle inequalities are derived with high probability. Since it turns out that DP suffers some deficiencies when estimating smooth functions, refinements involving smoothing of the residuals are introduced and analyzed. Theoretical bounds established in the fixed design setting under mild assumptions such as the boundedness of the kernel. When focusing on the smoothed discrepancy principle, such bounds are even extended to the random design setting by means of a new change-of-norm argument

Joint work with Markus Reiß(Humboldt) and Martin Wahl (Humboldt). This work has been already presented several times in seminars.

7.37. Axis 3: Short-term air temperature forecasting using Nonparametric Functional Data Analysis and SARMA models

Participant: Sophie Dabo-Niang

Air temperature is a significant meteorological variable that affects social activities and economic sectors. In this paper, a non-parametric and a parametric approach are used to forecast hourly air temperature up to 24 h in advance. The former is a regression model in the Functional Data Analysis framework. The nonlinear regression operator is estimated using a kernel function. The smoothing parameter is obtained by a cross-validation procedure and used for the selection of the optimal number of closest curves. The other method applied is a Seasonal Autoregressive Moving Average (SARMA) model, the order of which is determined by the Bayesian Information Criterion. The obtained forecasts are combined using weights calculated based on the forecast errors. The results show that SARMA has a better performance for the first 6 forecasted hours, after which the Non-Parametric Functional Data Analysis (NPFDA) model provides superior results. Forecast pooling improves the accuracy of the forecasts.

It is a joint work with Stelian Curceac (Rothamsted Research, UK) Camille Ternynck (CERIM, Université de Lille) Taha B.M.J. Ouarda (INRS, Québec, Canada) Fateh Chebana (INRS, Québec, Canada). This work has been published in the journal Environmental Modelling and Software [18].

7.38. Axis 3: Mathematical Modeling and Study of Random or Deterministic Phenomena

Participant: Sophie Dabo-Niang

In order to identify mathematical modeling (including functional data analysis) and interdisciplinary research issues in evolutionary biology, epidemiology, epistemology, environmental and social sciences encountered by researchers in Mayotte, the first international conference on mathematical modeling (CIMOM'18) was held in Dembéni, Mayotte, from November 15 to 17, 2018, at the Centre Universitaire de Formation et de Recherche. The objective was to focus on mathematical research with interdisciplinarity. This contribution is a book discusses key aspects of recent developments in applied mathematical analysis and modeling. It was written after the international conference on mathematical modeling in Mayotte, where a call for chapters of the book was made. They were written in the form of journal articles, with new results extending the talks given during the conference and were reviewed by independent reviewers and book publishers It highlights a wide range of applications in the fields of biological and environmental sciences, epidemiology and social perspectives. Each chapter examines selected research problems and presents a balanced mix of theory and applications on some selected topics. Particular emphasis is placed on presenting the fundamental developments in mathematical analysis and modeling and highlighting the latest developments in different fields of probability and statistics. The chapters are presented independently and contain enough references to allow the reader to explore the various topics presented.

It is a joint work with Solym Manou-Abi and Jean-Jacques Salone (University of Mayotte, France). This book is to appear Wiley (ISTE) [21].

7.39. Axis 3: Categorical functional data analysis

Participants: Cristian Preda, Quentin Grimonprez, Vincent Vandewalle.

The research on functional data analysis is very actual. The R package "fda" is the most famous one implementing methodology for functional data. To the best of our knowledge, and quite surprisingly, there is no recent researches devoted to categorical functional data despite its ability to model real situations in different fields of applications: health and medicine (status of a patient over time), economy (status of the market), sociology (evolution of social status), and so on. We have developed the methodology to visualize, do dimension reduction and extract feature from categorical functional data. For this, the *cfda* R package has been developed.

7.40. Axis 4: Proteomic signature of early death in heart failure patients

Participants: Guillemette Marot, Vincent Vandewalle.

Heart failure (HF) remains a main cause of mortality worldwide. Risk stratification of patients with systolic chronic HF is critical to identify those who may benefit from advanced HF therapies. The aim of this study is to identify plasmatic proteins that could predict the early death (within 3 years) of HF patients with reduced ejection fraction hospitalized in CHRU de Lille. In this framework, we have performed LASSO logistic regression to perform variable selection in order to select candidates protein to predict early death in HF patients. An article has been accepted in Scientific Reports [19].

This is a joint work with Marie Cuvelliez, Florence Pinet and Christophe Bauters from INSERM.

7.41. Axis 4: Statistical analysis of high-throughput proteomic data

Participants: Guillemette Marot, Vincent Vandewalle, Wilfried Heyse.

From March until August 2019, Guillemette Marot and Vincent Vandewalle have supervised the internship of Wilfried Heyse (Master 2 Ingénieurie de Sytèmes Numériques). The purpose of this internship was to identify new circulating biomarkers of left ventricular remodelling (LVR) in patients suffering from myocardial infarction (MI). The aim is to precisely identify earlier after MI the patients at high risk of developing LVR that is quantified by imaging one year after MI. For that purpose, high throughput proteomic approach was used. This technology allows the measurement of 5000 proteins simultaneously. In parallel to these measures corresponding to the concentration of a protein in a plasma sample collected from one patient at a specific time, echocardiographic and clinical information will be collected on each of the 200 patients. Several approaches

have been used to predict the LVR based on proteins measurements. In particular penalized regression such as LASSO and variable clustering. Wilfried Heyse has now started a Phd Thesis granted by INSERM and supervised by Christophe Bauters, Guillemette Marot and Vincent Vandewalle. One of the main challenge is to take into account the variations of the biomarkers according to the time (several measurement times), in order the improve the understanding of biological mechanisms involved on LVR.

This is a joint work with Florence Pinet and Christophe Bauters from INSERM.

7.42. Axis 4: Linking different kinds of Omics data through a model-based clustering approach

Participants: Guillemette Marot, Vincent Vandewalle, Wilfried Heyse.

In this work, a mixture model allowing for genes clustering using both microarray (continuous) and RNAseq (count) expression data is proposed. More generally, it answers the clustering of variables issue, when variables are of different kinds (continuous and discrete here). Variables describing the same gene are constrained to belong to the same cluster. This constraint allows us to obtain a model that links the microarray and RNAseq measurements without needing parametric constraints on the form of this link. The proposed approach has been illustrated on simulated data, as well as on real data from TCGA (The Cancer Genome Atlas). It has been presented in an international conference [49].

This is a joint work with Camille Ternynck from EA2694.

7.43. Axis 4: Real-time Audio Sources Classification

Participants: Christophe Biernacki, Maxime Baelde.

This work addresses the recurring challenge of real-time monophonic and polyphonic audio source classification. The whole power spectrum is directly involved in the proposed process, avoiding complex and hazardous traditional feature extraction. It is also a natural candidate for polyphonic events thanks to its additive property in such cases. The classification task is performed through a nonparametric kernel-based generative modeling of the power spectrum. Advantage of this model is twofold: it is almost hypothesis free and it allows to straightforwardly obtain the maximum a posteriori classification rule of online signals. Moreover it makes use of the monophonic dataset to build the polyphonic one. Then, to reach the real-time target, the complexity of the method can be tuned by using a standard hierarchical clustering preprocessing of sound models, revealing a particularly efficient computation time and classification accuracy trade-off. The proposed method reveals encouraging results both in monophonic and polyphonic classification tasks on benchmark and owned datasets, even in real-time situations. This method also has several advantages compared to the state-of-the-art methods include a reduced training time, no hyperparameters tuning, the ability to control the computation - accuracy trade-off and no training on already mixed sounds for polyphonic classification. This work is now published in an international journal [16] and Maxime Baelde defended his PhD thesis on this topic this year [11].

It is a joint work with Raphaël Greff, from the A-Volute company.

7.44. Axis 4: Matching of descriptors evolving over time

Participants: Christophe Biernacki, Anne-Lise Bedenel.

In the web domain, and in particular for insurance comparison, data constantly evolve, implying that it is difficult to directly exploit them. For example, to do a classification, performing standard learning processes require data descriptors equal for both learning and test samples. Indeed, for answering web surfer expectation, online forms whence data come from are regularly modified. So, features and data descriptors are also regularly modified. In this work, it is introduced a process to estimate and understand connections between transformed data descriptors. This estimated matching between descriptors will be a preliminary step before applying later classical learning methods. Anne-Lise Bedenel defended her PhD thesis on this topic this year [12].

It is a joint work with Laetitia Jourdan, from Université de Lille.

7.45. Axis 4: Supervised multivariate discretization and levels merging for logistic regression

Participants: Christophe Biernacki, Vincent Vandewalle, Adrien Ehrhardt.

For regulatory and interpretability reasons, the logistic regression is still widely used by financial institutions to learn the refunding probability of a loan given the applicants characteristics from historical data. Although logistic regression handles naturally both quantitative and qualitative data, three ad hoc pre-processing steps are usually performed: firstly, continuous features are discretized by assigning factor levels to predetermined intervals; secondly, qualitative features, if they take numerous values, are grouped; thirdly, interactions (products between two different features) are sparsely introduced. By reinterpreting these discretized (resp. grouped) features as latent variables and by modeling the conditional distribution of each of these latent variables given each original feature with a polytomous logistic link (resp. contingency table), a novel model-based resolution of the discretization problem is introduced. Estimation is performed via a Stochastic Expectation-Maximization (SEM) algorithm and a Gibbs sampler to find the best discretization (resp. grouping) scheme w.r.t. any classical logistic regression loss (AIC, BIC, test set AUC,...). For detecting interacting features, the same scheme is used by replacing the Gibbs sampler by a Metropolis-Hastings algorithm. The good performances of this approach are illustrated on simulated and real data from Credit Agricole Consumer Finance. Adrien Ehrhardt defended his PhD thesis on this topic this year [13]. A preprint is being finalized to be submitted to an international journal or conference [62].

This is a joint work with Philippe Heinrich from Université de Lille.

7.46. Axis 4: MASSICCC Platform for SaaS Software Availability

Participant: Christophe Biernacki.

MASSICCC is a demonstration platform giving access through a SaaS (service as a software) concept to data analysis libraries developed at Inria. It allows to obtain results either directly through a website specific display (specific and interactive visual outputs) or through an R data object download. It started in October 2015 for two years and is common to the Modal team (Inria Lille) and the Select team (Inria Saclay). In 2016, two packages have been integrated: Mixmod and MixtComp (see the specific section about MixtComp). In 2017, the BlockCluster package has been integrated and also a particular attention to provide meaningful graphical outputs (for Mixmod,MixtComp and BlockCluster) directly in the web platform itself has led to some specific developments. In 2019, MASSICCC has been presented to a workshop [39]. In addition, the Mixtcomp software is now available on the CRAN depository. Currently, a preprint for an international journal dedicated to software is also in progress.

The MASSICCC platform is available here in the web: https://massiccc.lille.inria.fr

7.47. Axis 4: Domain adaptation from a pre-trained source model

Participants: Christophe Biernacki, Pascal Germain, Luxin Zhang.

Traditional statistical learning paradigm assumes the consistency between train and test data distributions. This rarely holds in many real-life applications. The domain adaptation paradigm proposes a variety of techniques to overcome this issue. Most of the works in this area seek either for a latent space where source and target data share the same distribution, or for a transformation of the source distribution to match the target one. Both strategies require learning a model on the transformed source data. An original scenario is studied where one is given a model that has been constructed using expertise on the source data that is not accessible anymore. To use directly this model on target data, we propose to learn a transformation from the target domain to the source domain. Up to our knowledge, this is a new perspective on domain adaptation. This learning problem is introduced and formalized. We study the assumptions and the sufficient conditions mandatory to guarantee a good accuracy when using the source model directly on transformed target data. By pursuing this idea, a new domain adaptation method based on optimal transport is proposed. We experiment our method on a fraud detection problem.

Luxin Zhang begun his PhD thesis on this topic and presented this early result in an international conference [51].

It is a joint work with Yacine Kessaci, both from Worldline.

7.48. Axis 4: Reject Inference Methods in Credit Scoring: a rational review

Participants: Christophe Biernacki, Vincent Vandewalle, Adrien Ehrhardt.

The granting process of all credit institutions is based on the probability that the applicant will refund his/her loan given his/her characteristics. This probability also called score is learnt based on a dataset in which rejected applicants are de facto excluded. This implies that the population on which the score is used will be different from the learning population. Thus, this biased learning can have consequences on the scorecard's relevance. Many methods dubbed reject inference have been developed in order to try to exploit the data available from the rejected applicants to build the score. However most of these methods are considered from an empirical point of view, and there is some lack of formalization of the assumptions that are really made, and of the theoretical properties that can be expected. We propose a formalisation of these usually hidden assumptions for some of the most common reject inference methods, and we discuss the improvement that can be expected. These conclusions are illustrated on simulated data and on real data from Credit Agricole Consumer Finance (CACF), a major European loan issuer. Adrien Ehrhardt defended his PhD thesis on this topic this year [13]. A preprint is being finalized to be submitted to an international journal or conference.

This is a joint work with Philippe Heinrich from Université de Lille.

7.49. Other: Projection Under Pairwise Control

Participant: Christophe Biernacki.

Visualization of high-dimensional and possibly complex (non-continuous for instance) data onto a low-dimensional space may be difficult. Several projection methods have been already proposed for displaying such high-dimensional structures on a lower-dimensional space, but the information lost is not always easy to use. Here, a new projection paradigm is presented to describe a non-linear projection method that takes into account the projection quality of each projected point in the reduced space, this quality being directly available in the same scale as this reduced space. More specifically, this novel method allows a straightforward visualization data in R2 with a simple reading of the approximation quality, and provides then a novel variant of dimensionality reduction. This work is now under minor revision in an international journal [53].

It is a joint work with Hiba Alawieh and Nicolas Wicker, both from Université de Lille.

RAPSODI Project-Team

7. New Results

7.1. Modeling and numerical simulation of complex fluids

In [38], N. Peton, C. Cancès *et al.* propose a new water flow driven forward stratigraphic model. Stratigraphy is a discipline of physics that aims at predicting the geological composition of the subsoil. The model enjoys the following particularities. First, the water surface flow is modelled at the continuous level, in opposition to what is currently done in this community. Second, the model incorporates a constraint on the erosion rate. A stable numerical scheme is proposed to simulate the model.

In [14], A. Ait Hammou Oulhaj and D. Maltese adapt the (positive) nonlinear Control Volume Finite Element scheme of [83] to the simulation of seawater intrusion in the subsoil nearby coastal regions. The proposed scheme is convergent even if the porous medium is anisotropic.

In [25], [41], C. Cancès *et al.* study an original model of degenerate Cahn–Hilliard type. Similarly to the classical degenerate Cahn–Hilliard model, the model can be interpreted as the gradient flow of a Ginzburg–Landau type energy, but the geometry considered here allows for more flexibility and the system thus dissipates faster than the usual degenerate Cahn–Hilliard system. Numerical evidences of this fact are given. Then, the existence of a solution to the model is established thanks to the convergence of a minimizing movement scheme.

In [19], I. Lacroix-Violet *et al.* generalize to the Navier–Stokes–Korteweg (with density-dependent viscosities satisfying the BD relation) and Euler–Korteweg systems a recent relative entropy proposed in [80]. As a concrete application, this helps justifying mathematically the convergence between global weak solutions of the quantum Navier–Stokes system and dissipative solutions of the quantum Euler system when the viscosity coefficient tends to zero. The results are based on the fact that Euler–Korteweg systems and corresponding Navier–Stokes–Korteweg systems can be reformulated through an augmented system. As a by-product of the analysis, I. Lacroix-Violet *et al.* show that this augmented formulation helps to define relative entropy estimates for the Euler–Korteweg systems in a simpler way and with less hypotheses compared to recent works [97], [100].

In [22], C. Calgaro, C. Colin, E. Creusé *et al.* investigate a specific low-Mach model for which the dynamic viscosity of the fluid is a specific function of the density. The model is reformulated in terms of the temperature and velocity, with nonlinear temperature equation, and strong solutions are considered. In addition to a local-in-time existence result for strong solutions, some convergence rates of the error between the approximation and the exact solution are obtained, following the same approach as Guillén-González *et al.* [103], [104].

In [21], C. Calgaro, C. Colin, and E. Creusé derive a combined Finite Volume-Finite Element scheme for a low-Mach model, in which a temperature field obeying an energy law is taken into account. The continuity equation is solved, whereas the state equation linking temperature, density, and thermodynamic pressure is imposed implicitly. Since the velocity field is not divergence-free, the projection method solving the momentum equation has to be adapted. This combined scheme preserves some steady-states, and ensures a discrete maximum principle on the density. Numerical results are provided and compared to other approaches using purely Finite Element schemes, on a benchmark consisting in particular in a transient injection flow [74], [101], [69], as well as in the natural convection of a flow in a cavity [108], [105], [101], [69].

In [20], C. Calgaro, C. Colin, and E. Creusé propose a combined Finite Volume-Finite Element scheme for the solution of a specific low-Mach model expressed in the velocity, pressure and temperature variables. The dynamic viscosity of the fluid is given by an explicit function of the temperature, leading to the presence of a so-called Joule term in the mass conservation equation. First, they prove a discrete maximum principle for the temperature. Second, the numerical fluxes defined for the Finite Volume computation of the temperature are efficiently derived from the discrete Finite Element velocity field obtained by the solution of the momentum

equation. Several numerical tests are presented to illustrate the theoretical results and to underline the efficiency of the scheme in terms of convergence rates.

In [46], C. Calgaro and E. Creusé introduce a Finite Volume method to approximate the solution of a convection-diffusion equation involving a Joule term. They propose a way to discretize this so-called "Joule effect" term in a consistent manner with respect to the nonlinear diffusion one, in order to ensure some maximum principle properties on the solution. They investigate the numerical behavior of the scheme on two original benchmarks.

7.2. Numerical simulation in low-frequency electromagnetism

In [32], [31], E. Creusé *et al.* investigate the behavior of some Finite Element error estimators in the context of low-frequency elecromagnetism simulations, to underline the main differences in some practical situations. A more theoretical contribution is also developed to prove the equivalence of some usual discrete gauge conditions. Once again, their numerical behaviors are compared on some characteristic benchmarks.

In [58], F. Chave, S. Lemaire *et al.* introduce a three-dimensional Hybrid High-Order (HHO) method for magnetostatic problems. The proposed method is easy to implement, supports general polyhedral meshes, and allows for arbitrary orders of approximation.

7.3. Structure-preserving numerical methods

In [54], C. Cancès *et al.* propose a Finite Element scheme for the numerical approximation of degenerate parabolic problems in the form of a nonlinear anisotropic Fokker–Planck equation. The scheme is energy-stable, only involves physically motivated quantities in its definition, and is able to handle general unstructured grids. Its convergence is rigorously proven thanks to compactness arguments, under very general assumptions. Although the scheme is based on Lagrange Finite Elements of degree 1, it is locally conservative after a local post-processing giving rise to an equilibrated flux. This also allows to derive a guaranteed *a posteriori* error estimate for the approximate solution. Numerical experiments are presented in order to give evidence of a very good behavior of the proposed scheme in various situations involving strong anisotropy and drift terms.

In [55], C. Chainais-Hillairet and M. Herda apply an iterative energy method à la de Giorgi in order to establish L^∞ bounds for numerical solutions of noncoercive convection-diffusion equations with mixed Dirichlet-Neumann boundary conditions.

In [23], C. Cancès, C. Chainais-Hillairet *et al.* study a finite volume scheme for a degenerate cross-diffusion system describing the ion transport through biological membranes. The strongly coupled equations for the ion concentrations include drift terms involving the electric potential, which is coupled to the concentrations through the Poisson equation. The finite volume scheme is based on two-point flux approximations with "double" upwind mobilities. The existence of solutions to the fully discrete scheme is proven. When the particles are not distinguishable and the dynamics is driven by cross-diffusion only, it is shown that the scheme preserves the structure of the equations like nonnegativity, upper bounds, and entropy dissipation.

In [51], C. Cancès and B. Gaudeul propose a two-point flux approximation finite volume scheme for the approximation of the solutions to an entropy dissipative cross-diffusion system. The scheme is shown to preserve several key properties of the continuous system, among which positivity and decay of the entropy. Numerical experiments illustrate the behavior of the scheme.

In [48], C. Cancès, C. Chainais-Hillairet, B. Gaudeul *et al.* consider an unipolar degenerate drift-diffusion system arising in the modeling of organic semiconductors. They design four different finite volume schemes based on four different formulations of the fluxes. They provide a stability analysis and existence results for the four schemes; the convergence is established for two of them.

In [24], C. Cancès *et al.* compare energy-stable finite volume schemes for multiphase flows in porous media with schemes based on the Wasserstein gradient flow structure of the equations, that has recently been highlighted in [3]. The model is approximated by means of the minimizing movement (or JKO) scheme, that C. Cancès *et al.* solve thanks to the ALG2-JKO scheme proposed in [76].

In [50], C. Cancès *et al.* propose a variational finite volume scheme for the computation of Wasserstein gradient flows. The discrete solution is the minimizer of a discrete action, keeping track at the discrete level of the optimal character of the gradient flow. The spatial discretization relies on upstream mobility fluxes, while an implicit linearization of the Wasserstein distance is used in order to reduce the computational cost by avoiding an inner time-stepping as in the related contributions of the literature.

In [61], T. Rey *et al.* present a new finite volume method for computing numerical approximations of a system of nonlocal transport equations modeling interacting species. In this work, the nonlocal continuity equations are treated as conservative transport equations with a nonlocal, nonlinear, rough velocity field. Some properties of the method are analyzed, and numerical simulations are performed.

In [15], I. Lacroix-Violet *et al.* are interested in the numerical integration in time of nonlinear Schrödinger equations using different methods preserving the energy or a discrete analog of it. In particular, they give a rigorous proof of the order of the relaxation method (presented in [78] for cubic nonlinearities) and they propose a generalized version that allows to deal with general power law nonlinearities. Numerical simulations for different physical models show the efficiency of these methods.

7.4. Cost reduction for numerical methods

In [36], S. Lemaire builds a bridge between the Hybrid High-Order [93] and Virtual Element [75] methods, which are the two main new-generation approaches to the arbitrary-order approximation of PDEs on meshes with general, polytopal cells. The Virtual Element method writes in functional terms and is naturally conforming; at the opposite, the Hybrid High-Order method writes in algebraic terms and is naturally nonconforming. It has been remarked a few years ago that the Hybrid High-Order method can be viewed as a nonconforming version of the Virtual Element method. Here, S. Lemaire ends up unifying the Hybrid High-Order and Virtual Element approaches by showing that the Virtual Element method can be reformulated as a (newborn) conforming Hybrid High-Order method. This parallel has interesting consequences as it sheds new light on the *a priori* analysis of Virtual Element methods, and on the differences between the conforming and nonconforming cases.

In [30], [40], S. Lemaire *et al.* design and analyze (in the periodic setting) arbitrary-order nonconforming multiscale methods for highly oscillatory elliptic problems, which are applicable on coarse grids that may feature general polytopal cells. The construction of these methods is based on the Hybrid High-Order framework [93]. As standard with such multiscale approaches, the general workflow of the method splits into an offline, massively parallelizable stage where all fine-scale computations are performed, and the online, fully coarse-scale stage.

In [52], C. Cancès and D. Maltese propose a reduced model for the migration of hydrocarbons in heterogeneous porous media. Their model keeps track of the time variable. This allows to compute steady-states that cannot be reached by the commonly used ray-tracing and invasion-percolation algorithms. An efficient finite volume scheme allowing for very large time steps is then proposed.

In [57], F. Chave proposes a new definition of the normal fracture diffusion-dispersion coefficient for a reduced model of passive transport in fractured porous media, and numerically studies the impact on the discrete solution on a few test-cases.

In [37], T. Rey *et al.* present high-order, fully explicit time integrators for nonlinear collisional kinetic equations, including the full Boltzmann equation. The methods, called projective integration, first take a few small steps with a simple, explicit method (forward Euler) to damp out the stiff components of the solution. Then, the time derivative is estimated and used in a Runge–Kutta method of arbitrary order. The procedure can be recursively repeated on a hierarchy of projective levels to construct telescopic projective integration methods. The method is illustrated with numerical results in one and two space dimensions.

In [60], I. Lacroix-Violet *et al.* introduce a new class of numerical methods for the time integration of evolution equations set as Cauchy problems of ODEs or PDEs. The systematic design of these methods mixes the Runge–Kutta collocation formalism with collocation techniques, in such a way that the methods are linearly implicit and have high order. The fact that these methods are implicit allows to avoid CFL conditions when

the large systems to integrate come from the space discretization of evolution PDEs. Moreover, these methods are expected to be efficient since they only require to solve one linear system of equations at each time step, and efficient techniques from the literature can be used to do so.

7.5. Asymptotic analysis

In [18], C. Cancès *et al.* derive the porous medium equation as the hydrodynamic limit of an interacting particle system which belongs to the family of exclusion processes, with nearest neighbor exchanges. The particles follow a degenerate dynamics, in the sense that the jump rates can vanish for certain configurations, and there exist blocked configurations that cannot evolve. Our approach, which is based on the relative entropy method, is tailored to deal with vanishing initial densities.

In [13], A. Ait Hammou Oulhaj, C. Cancès, C. Chainais-Hillairet *et al.* study the large-time behavior of the solutions to a two-phase extension of the porous media equation, which models the seawater intrusion problem. Their goal is to identify the self-similar solutions that correspond to steady-states of a rescaled version of the problem. They fully characterize the unique steady-states that are identified as minimizers of a convex energy and shown to be radially symmetric. Moreover, they prove the convergence of the solution to the time-dependent model towards the unique stationary state as time goes to infinity. They also provide numerical illustrations of the stationary states and exhibit numerical convergence rates.

In [16], C. Chainais-Hillairet *et al.* propose a new proof of existence of a solution to the scheme introduced in [1] for drift-diffusion systems, which does not require any assumption on the time step. The result relies on the application of a topological degree argument which is based on the positivity and on uniform-in-time upper bounds of the approximate densities. They also establish uniform-in-time lower bounds satisfied by the approximate densities. These uniform-in-time upper and lower bounds ensure the exponential decay of the scheme towards the thermal equilibrium as shown in [1].

In [26], C. Chainais-Hillairet and M. Herda study the large-time behavior of the solutions to Finite Volume discretizations of convection-diffusion equations or systems endowed with non-homogeneous Dirichlet and Neumann type boundary conditions. Their results concern various linear and nonlinear models such as Fokker–Planck equations, porous media equations, or drift-diffusion systems for semiconductors. For all of these models, some relative entropy principle is satisfied and implies exponential decay to the stationary state. They show that in the framework of Finite Volume schemes on orthogonal meshes, a large class of two-point monotone fluxes preserve this exponential decay of the discrete solution to the discrete steady-state of the scheme.

In [49], C. Cancès, C. Chainais-Hillairet, M. Herda *et al.* analyze the large-time behavior of a family of nonlinear finite volume schemes for anisotropic convection-diffusion equations set in a bounded bidimensional domain and endowed with either Dirichlet and/or no-flux boundary conditions. They show that the solutions to the two-point flux approximation (TPFA) and discrete duality finite volume (DDFV) schemes under consideration converge exponentially fast toward their steady-state. The analysis relies on discrete entropy estimates and discrete functional inequalities. As a by-product of their analysis, they establish new discrete Poincaré–Wirtinger, Beckner and logarithmic Sobolev inequalities. Their theoretical results are illustrated by numerical simulations.

In [56], C. Chainais-Hillairet *et al.* introduce a nonlinear DDFV scheme for an anisotropic linear convection-diffusion equation with mixed boundary conditions and establish the exponential decay of the scheme towards its steady-state.

In [39], A. Zurek studies the large-time regime of the moving interface appearing in a concrete carbonation model. He proves that the approximate free boundary, given by an implicit-in-time Finite Volume scheme, propagates in time following a $\sqrt{t}$-law. This result is illustrated by numerical experiments.

In [17], M. Herda, T. Rey *et al.* are interested in the asymptotic analysis of a Finite Volume scheme for one-dimensional linear kinetic equations, with either Fokker–Planck or linearized BGK collision operator. Thanks to appropriate uniform estimates, they establish that the proposed scheme is asymptotic-preserving in the diffusive limit. Moreover, they adapt to the discrete framework the hypocoercivity method proposed in [95]

to prove the exponential return to equilibrium of the approximate solution. They obtain decay estimates that are uniform in the diffusive limit. Finally, they present an efficient implementation of the proposed numerical schemes, and perform numerous numerical simulations assessing their accuracy and efficiency in capturing the correct asymptotic behaviors of the models.

In [44], M. Herda *et al.* are interested in the large-time behavior of linear kinetic equations with heavy-tailed local equilibria. Their main contribution concerns the kinetic Lévy–Fokker–Planck equation, for which they adapt hypocoercivity techniques in order to show that solutions converge exponentially fast to the global equilibrium. Compared to the classical kinetic Fokker–Planck equation, the issues here concern the lack of symmetry of the non-local Lévy–Fokker–Planck operator and the understanding of its regularization properties. As a complementary related result, they also treat the case of the heavy-tailed BGK equation.

In [35], M. Herda *et al.* consider various sets of Vlasov–Fokker–Planck equations modeling the dynamics of charged particles in a plasma under the effect of a strong magnetic field. For each of them, in a regime where the strength of the magnetic field is effectively stronger than that of collisions, they first formally derive asymptotically reduced models. In this regime, strong anisotropic phenomena occur; while equilibrium along magnetic field lines is asymptotically reached, the asymptotic models capture a nontrivial dynamics in the perpendicular directions. They do check that in any case the obtained asymptotic model defines a well-posed dynamical system and when self-consistent electric fields are neglected they provide a rigorous mathematical justification of the formally derived systems. In this last step they provide a complete control on solutions by developing anisotropic hypocoercive estimates.

In [45], T. Rey *et al.* propose a new mathematical model intended to describe dynamically the evolution of knowledge in structured societies of interacting individuals. This process, termed cumulative culture, has been extensively studied by evolutionary anthropologists, both theoretically and experimentally. Some of the mathematical properties of the new model are analyzed, and exponential convergence towards a global equilibrium is shown for a simplified model. A numerical method is finally proposed to simulate the complete model.

In [43], following the ideas of V. V. Zhikov and A. L. Pyatnitskii, and more precisely the stochastic two-scale convergence, B. Merlet *et al.* establish a homogenization theorem in a stochastic setting for two nonlinear equations: the equation of harmonic maps into the sphere and the Landau–Lifshitz equation. Homogenization results for nonlinear problems are known to be difficult. In this particular case the equations have strong nonlinear features, in particular, in general their solutions are not unique. Here the authors take advantage of the different equivalent definitions of weak solutions to the nonlinear problem to apply typical linear homogenization recipes.

7.6. Applied calculus of variations

In [34], B. Merlet *et al.* study a variational problem which models the behavior of topological singularities on the surface of a biological membrane in P_β-phase (see [112]). The problem combines features of the Ginzburg–Landau model in 2D and of the Mumford–Shah functional. As in the classical Ginzburg–Landau theory, a prescribed number of point vortices appear in the moderate energy regime; the model allows for discontinuities, and the energy penalizes their length. The novel phenomenon here is that the vortices have a fractional degree $1/m$ with m prescribed. Those vortices must be connected by line discontinuities to form clusters of total integer degrees. The vortices and line discontinuities are therefore coupled through a topological constraint. As in the Ginzburg–Landau model, the energy is parameterized by a small length scale $\varepsilon > 0$. B. Merlet *et al.* perform a complete Γ-convergence analysis of the model as $\varepsilon \downarrow 0$ in the moderate energy regime. Then, they study the structure of minimizers of the limit problem. In particular, the line discontinuities of a minimizer solve a variant of the Steiner problem.

In [27], B. Merlet *et al.* consider the branched transportation problem in 2D associated with a cost per unit length of the form $1 + \beta\theta$ where θ denotes the amount of transported mass and $\beta > 0$ is a fixed parameter (notice that the limit case $\beta = 0$ corresponds to the classical Steiner problem). Motivated by the numerical approximation of this problem, they introduce a family of functionals $(\{\mathcal{F}_\varepsilon\}_{\varepsilon>0})$ which approximate the above

branched transport energy. They justify rigorously the approximation by establishing the equicoercivity and the Γ-convergence of $\{\mathcal{F}_\varepsilon\}$ as $\varepsilon \downarrow 0$. The functionals are modeled on the Ambrosio–Tortorelli functional and are easy to optimize in practice. Numerical evidences of the efficiency of the method are presented.

In [28], B. Merlet *et al.* establish new results on the approximation of k-dimensional surfaces (k-rectifiable currents) by polyhedral surfaces with convergence in h-mass and with preservation of the boundary (the approximating polyhedral surface has the same boundary as the limit). This approximation result is required in the convergence study of [29].

In [29], B. Merlet *et al.* consider a generalization of branched transportation in arbitrary dimension and codimension: minimize the h-mass of some oriented k-dimensional branched surface in $\mathbf{R}^n$ with some prescribed boundary. Attached to the surface is a multiplicity $m(x)$ which is not necessarily an integer and is a conserved quantity (Kirchhoff current law is satisfied at branched points). The h-mass is defined as the integral of a cost $h(|m(x)|)$ over the branched surface. As usual in branched transportation, the cost function is a lower-semicontious, sublinear increasing function with $h(0) = 0$ (for instance $h(m) = \sqrt{1 + am^2}$ if $m \neq 0$ and $h(0) = 0$). For numerical purpose, it is convenient to approximate the measure defined by the k-dimensional surfaces by smooth functions in $\mathbf{R}^n$. In this spirit, B. Merlet *et al.* propose phase field approximations of the branched surfaces and of their energy in the spirit of the Ambrosio–Tortorelli functional. The convergence of these approximations towards the original k-dimensional branched transportation problem is established in the sense of Γ-convergence. Next, considering the cost $h(m) = \sqrt{1 + am^2}$ and sending a to 0, a phase field approximation of the Plateau problem is obtained. Numerical experiments show the efficiency of the method. These numerical results are exceptional as they are obtained without any guess on the topology of the minimizing k-surface (as opposed to methods based on parameterizations of the k-surface).

In [33], [62], B. Merlet *et al.* study a family of functionals penalizing oblique oscillations. These functionals naturally appear in some variational problems related to pattern formation and are somewhat reminiscent of those introduced by Bourgain, Brezis and Mironescu to characterize Sobolev functions. More precisely, for a function u defined on a tensor product $\Omega_1 \times \Omega_2$, the family of functionals $\{E_\varepsilon(u)\}_{\varepsilon>0}$ that we consider vanishes if u is of the form $u(x_1)$ or $u(x_2)$. We prove the converse property and related quantitative results. In particular, we describe the fine properties of functions with $\sup_\varepsilon E_\varepsilon(u) < \infty$ by showing that roughly, such u is piecewise of the form $u(x_1)$ or $u(x_2)$ on domains separated by lines where the energy concentrates. It turns out that this problem naturally leads to the study of various differential inclusions and has connections with branched transportation models.

7.7. Approximation theory

In [59], M. Herda *et al.* propose an iterative algorithm for the numerical computation of sums of squares of polynomials approximating given data at prescribed interpolation points. The method is based on the definition of a convex functional G arising from the dualization of a quadratic regression over the Cholesky factors of the sum of squares decomposition. In order to justify the construction, the domain of G, the boundary of the domain and the behavior at infinity are analyzed in details. When the data interpolate a positive univariate polynomial, we show that in the context of the Lukacs sum of squares representation, G is coercive and strictly convex which yields a unique critical point and a corresponding decomposition in sum of squares. For multivariate polynomials which admit a decomposition in sum of squares and up to a small perturbation of size ε, G^ε is always coercive and so its minimum yields an approximate decomposition in sum of squares. Various unconstrained descent algorithms are proposed to minimize G. Numerical examples are provided, for univariate and bivariate polynomials.

In [47], M. Herda *et al.* investigate the numerical approximation of bounded functions by polynomials satisfying the same bounds. The contribution makes use of the recent algebraic characterization found in [91] and [92] where an interpretation of monovariate polynomials with two bounds is provided in terms of a quaternion algebra and the Euler four-squares formulas. Thanks to this structure, the authors generate a new nonlinear projection algorithm onto the set of polynomials with two bounds. The numerical analysis of the method provides theoretical error estimates showing stability and continuity of the projection. Some numerical tests illustrate this novel algorithm for constrained polynomial approximation.

RMOD Project-Team

7. New Results

7.1. Dynamic Languages: Virtual Machines

Illicium A modular transpilation toolchain from Pharo to C. The Pharo programming language runs on the OpenSmalltalk-VM. This Virtual Machine (VM) is mainly written in Slang, a subset of the Smalltalk language dedicated to VM development. Slang is transpiled to C using the Slang-to-C transpiler. The generated C is then compiled to produce the VM executable binary code. Slang is a powerful dialect for generating C because it benefits from the tools of the Smalltalk environment, including a simulator that runs and debugs the VM. However, the Slang-to-C transpiler is often too permissive. For example, the Slang-to-C transpiler generates invalid C code from some Smalltalk concepts it does not support. This makes the Slang code hard to debug as the errors are caught very late during the development process, which is worsen by the loss of the mapping between the generated C code and Slang. The Slang-to-C transpiler is also hard to extend or adapt to modify part of the translation process. We present Illicium, a new modular transpilation toolchain based on a subset of Pharo targeting C through AST transformations. This toolchain translates the Pharo AST into a C AST to generate C code. Using ASTs as source and target artifacts enables analysis, modification and validation at different levels during the translation process. The main translator is split into smaller and replaceable translators to increase modularity. Illicium also allows the possibility to introduce new translators and to chain them together, increasing reusability. To evaluate our approach, we show with a use case how to extend the transpilation process with a translation that requires changes not considered in the original C AST. [7]

GildaVM: a Non-Blocking I/O Architecture for the Cog VM. The OpenSmalltalk virtual machine (VM) was historically designed as a single-threaded VM. All VM code including the Smalltalk interpreter, the garbage collector and the just-in-time compiler run in the same single native thread. While this VM provides concurrency through green threads, it cannot take advantage of multi-core processors. This architecture performs really well in practice until the VM accesses external resources such as e.g., FFI callouts, which block the single VM thread and prevent green threads to benefit from the processor. We present GildaVM, a multi-threaded VM architecture where one thread at a time executes the VM while allowing non-blocking I/O in parallel. The ownership of the VM is orchestrated by a Global Interpreter Lock (GIL) as in the standard implementations of Python and Ruby. However, within a single VM thread concurrency is still possible through green threads. We present a prototype implementation of this architecture running on top of the Stack flavour of the OpenSmalltalk VM. We finally evaluate several aspects of this architecture like FFI and thread-switch overhead. While current benchmarks show good results for long FFI calls, short FFI calls require more research to minimize the overhead of thread-switch. [9]

7.2. Dynamic Languages: Language Constructs for Modular Design

Magic Literals in Pharo Literals are constant values (numbers, strings, etc.) used in the source code. Magic literals are the ones used without a clear explanation of their meaning. Presence of such literals harms source code readability, decreases its modularity, and encourages code duplication. Identifying magic literals is not straightforward. A literal can be considered self-explanatory in one context and magic in another. We need a heuristic to help developers spot magic literals. We study and characterize the literals in Pharo. We implemented a heuristic to detect magic literals and integrated it as a code critic rule for System Browser and Critics Browser in Pharo 7. We run our heuristic on 112,500 Pharo methods which reported 23,292 magic literals spread across 8,986 methods. We manually validated our approach on a random subset of 100 methods and found that 62% of the reported literals in those methods are indeed magic. [3]

Towards easy program migration using language virtualization Migrating programs between language versions is a daunting task. A developer writes a program in a particular version of a language and cannot foresee future language changes. In this article, we explore a solution to gradual program migration based on virtualization at the programming language level. Our language virtualization approach adds a backwards-compatibility layer on top of a recent language version, allowing developers to load and run old programs on the more recent infrastructure. Developers are then able to migrate the program to the new language version or are able to run it as it is. Our virtualization technique is based on a dynamic module implementation and code intercession techniques. Migrated and non-migrated parts co-exist in the meantime allowing an incremental migration procedure. We validate it by migrating legacy Pharo programs, MuTalk and Fuel. [10]

7.3. Dynamic Languages: Debugging

Sindarin: A Versatile Scripting API for the Pharo Debugger Debugging is one of the most important and time consuming activities in software maintenance, yet mainstream debuggers are not well-adapted to several debugging scenarios. This has led to the research of new techniques covering specific families of complex bugs. Notably, recent research proposes to empower developers with scripting DSLs, plugin-based and moldable debuggers. However, these solutions are tailored to specific use-cases, or too costly for one-time-use scenarios. We argue that exposing a debugging scripting interface in mainstream debuggers helps in solving many challenging debugging scenarios. For this purpose, we present Sindarin, a scripting API that eases the expression and automation of different strategies developers pursue during their debugging sessions. Sindarin provides a GDB-like API, augmented with AST-bytecode-source code mappings and object-centric capabilities. To demonstrate the versatility of Sindarin, we reproduce several advanced breakpoints and non-trivial debugging mechanisms from the literature. [4]

Challenges in Debugging Bootstraps of Reflective Kernels The current explosion of embedded systems (i.e., IoT, Edge Computing) implies the need for generating tailored and customized software for these systems. Instead of using specific runtimes (e.g., MicroPython, eLua, mRuby), we advocate that bootstrapping specific language kernels is a promising higher-level approach because the process takes advantage of the generated language abstractions, easing the task for a language developer. Nevertheless, bootstrapping language kernels is still challenging because current debugging tools are not suitable for fixing the possible failures that occur during the process. We take the Pharo bootstrap process as an example to analyse the different challenges a language developer faces. We propose a taxonomy of failures appearing during bootstrap and their causes. Based on this analysis, we identify future research directions: (1) prevention measures based on the reification of implicit virtual machine contracts, and (2) hybrid debugging tools that unify the debugging of high-level code from the bootstrapped language with low-level code from the virtual machine. [6]

7.4. Software Reengineering

Decomposing God Classes at Siemens A group of developers at Siemens Digital Industry Division approached our team to help them restructure a large legacy system. Several problems were identified, including the presence of God classes (big classes with thousands of lines of code and hundred of methods). They had tried different approaches considering the dependencies between the classes, but none were satisfactory. Through interaction during the last three years with a lead software architect of the project, we designed a software visualization tool and an accompanying process that allows her to propose a decomposition of a God Class in a matter of one or two hours even without prior knowledge of the class (although actually implementing the decomposition in the source code could take a week of work). We present the process that was formalized to decompose God Classes and the tool that was designed. We give details on the system itself and some of the classes that were decomposed. The presented process and visualisations have been successfully used for the last three years on a real industrial system at Siemens. [1]

Rotten Green Tests Unit tests are a tenant of agile programming methodologies, and are widely used to improve code quality and prevent code regression. A green (passing) test is usually taken as a robust sign that the code under test is valid. However, some green tests contain assertions that are never executed. We call such tests Rotten Green Tests. Rotten Green Tests represent a case worse than a broken test: they report that

the code under test is valid, but in fact do not test that validity. We describe an approach to identify rotten green tests by combining simple static and dynamic call-site analyses. Our approach takes into account test helper methods, inherited helpers, and trait compositions, and has been implemented in a tool called DrTest. DrTest reports no false negatives, yet it still reports some false positives due to conditional use or multiple test contexts. Using DrTest we conducted an empirical evaluation of 19,905 real test cases in mature projects of the Pharo ecosystem. The results of the evaluation show that the tool is effective; it detected 294 tests as rotten-green tests that contain assertions that are not executed. Some rotten tests have been "sleeping" in Pharo for at least 5 years. [2]

Migrating GWT to Angular 6 using MDE In the context of a collaboration with Berger-Levrault, a major IT company, we are working on the migration of a GWT application to Angular. We focus on the GUI aspect of this migration which, even if both are web frameworks, is made difficult because they use different programming languages (Java for one, Typescript for the other) and different organization schemas (e.g. different XML files). Moreover, the new application must mimic closely the visual aspect of the old one so that the users of the application are not disturbed. We propose an approach in three steps that uses a meta-model to represent the GUI at a high abstraction level. We evaluated this approach on an application comprising 470 Java (GWT) classes representing 56 screens. We are able to model all the web pages of the application and 93% of the wid-gets they contain, and we successfully migrated (i.e., the result is visually equal to the original) 26 out of 39 pages (66%). We give examples of the migrated pages, both successful and not. [14] [11] [12]

Empirical Study of Programming to an Interface A popular recommendation to programmers in object-oriented software is to *program to an interface, not an implementation* (PTI). Expected benefits include increased simplicity from abstraction, decreased dependency on implementations, and higher flexibility. Yet, interfaces must be immutable, excessive class hierarchies can be a form of complexity, and *speculative generality* is a known code smell. To advance the empirical knowledge of PTI, we conducted an empirical investigation that involves 126 Java projects on GitHub, aiming to measuring the decreased dependency benefits (in terms of cochange). [13]

Exposing Test Analysis Results with DrTests Tests are getting the cornerstone of continuous development process and software evolution. Tests are the new gold. To improve test quality, a plethora of analyses is proposed such as test smells, mutation testing, test coverage. The problem is that each analysis often needs a particular way to expose its results to the developer. There is a need for an architecture supporting test running and analysis in a modular and extensible way. We present an extensible plugin-based architecture to run and report test results. DrTests is a new test browser that implements such plugin-based architecture. DrTests supports the execution of rotten tests, comments to tests, coverage and profiling tests. [5]

7.5. Blockchain Software Engineering

SmartAnvil: Open-Source Tool Suite for Smart Contract Analysis Smart contracts are new computational units with special properties: they act as classes with aspectual concerns; their memory structure is more complex than mere objects; they are obscure in the sense that once deployed it is difficult to access their internal state; they reside in an append-only chain. There is a need to support the building of new generation tools to help developers. Such support should tackle several important aspects: (1) the static structure of the contract, (2) the object nature of published contracts, and (3) the overall data chain composed of blocks and transactions. We present SmartAnvil an open platform to build software analysis tools around smart contracts. We illustrate the general components and we focus on three important aspects: support for static analysis of Solidity smart contracts, deployed smart contract binary analysis through inspection, and blockchain navigation and querying. SmartAnvil is open-source and supports a bridge to the Moose data and software analysis platform. [18]

The Influence Factors on Ethereum Transaction Fees In Ethereum blockchain, the user needs to set a Gas price to get a transaction processed and approved by Miners. To have the transaction executed, the Gas price has to be greater than or equal to the lowest Ethereum transaction fees. We present a set of data sampled every 15 seconds, from December 1st, 2018 to December 15, 2018, coming from different blockchain web APIs. The aim is to investigate whether and to what extent different variables - such as the number of pending

transactions, the value of the USD/Ether pair, average electricity prices around the world, and the number of miners - influence the Ethereum transaction fees. This study is relevant from an economic perspective because more and more companies in different economic fields are adopting Ethereum blockchain. From historical data analysis, we found that only some of these variables do have an influence. For example, the number of pending transactions and the number of miners have a major influence on Ethereum transaction fees when compared to the other variables. [8]

SEQUEL Project-Team

7. New Results

7.1. Decision-making Under Uncertainty

7.1.1. Reinforcement Learning

Model-Based Reinforcement Learning Exploiting State-Action Equivalence, [16]

Leveraging an equivalence property in the state-space of a Markov Decision Process (MDP) has been investigated in several studies. This paper studies equivalence structure in the reinforcement learning (RL) setup, where transition distributions are no longer assumed to be known. We present a notion of similarity between transition probabilities of various state-action pairs of an MDP, which naturally defines an equivalence structure in the state-action space. We present equivalence-aware confidence sets for the case where the learner knows the underlying structure in advance. These sets are provably smaller than their corresponding equivalence-oblivious counterparts. In the more challenging case of an unknown equivalence structure, we present an algorithm called ApproxEquivalence that seeks to find an (approximate) equivalence structure, and define confidence sets using the approximate equivalence. To illustrate the efficacy of the presented confidence sets, we present C-UCRL, as a natural modification of UCRL2 for RL in undiscounted MDPs. In the case of a known equivalence structure, we show that C-UCRL improves over UCRL2 in terms of regret by a factor of SA/C, in any communicating MDP with S states, A actions, and C classes, which corresponds to a massive improvement when C SA. To the best of our knowledge, this is the first work providing regret bounds for RL when an equivalence structure in the MDP is efficiently exploited. In the case of an unknown equivalence structure, we show through numerical experiments that C-UCRL combined with ApproxEquivalence outperforms UCRL2 in ergodic MDPs.

Practical Open-Loop Optimistic Planning, [25]

We consider the problem of online planning in a Markov Decision Process when given only access to a generative model, restricted to open-loop policies-i.e. sequences of actions-and under budget constraint. In this setting, the Open-Loop Optimistic Planning (OLOP) algorithm enjoys good theoretical guarantees but is overly conservative in practice, as we show in numerical experiments. We propose a modified version of the algorithm with tighter upper-confidence bounds, KL-OLOP, that leads to better practical performances while retaining the sample complexity bound. Finally, we propose an efficient implementation that significantly improves the time complexity of both algorithms.

Budgeted Reinforcement Learning in Continuous State Space, [20]

A Budgeted Markov Decision Process (BMDP) is an extension of a Markov Decision Process to critical applications requiring safety constraints. It relies on a notion of risk implemented in the shape of a cost signal constrained to lie below an-adjustable-threshold. So far, BMDPs could only be solved in the case of finite state spaces with known dynamics. This work extends the state-of-the-art to continuous spaces environments and unknown dynamics. We show that the solution to a BMDP is a fixed point of a novel Budgeted Bellman Optimality operator. This observation allows us to introduce natural extensions of Deep Reinforcement Learning algorithms to address large-scale BMDPs. We validate our approach on two simulated applications: spoken dialogue and autonomous driving.

Regret Bounds for Learning State Representations in Reinforcement Learning, [29]

We consider the problem of online reinforcement learning when several state representations (mapping histories to a discrete state space) are available to the learning agent. At least one of these representations is assumed to induce a Markov decision process (MDP), and the performance of the agent is measured in terms of cumulative regret against the optimal policy giving the highest average reward in this MDP representation. We propose an algorithm (UCB-MS) with O($\sqrt{T}$) regret in any communicating MDP. The regret bound shows that UCB-MS automatically adapts to the Markov model and improves over the currently known best bound of order O(T 2/3).

Planning in entropy-regularized Markov decision processes and games, [24]

We propose SmoothCruiser, a new planning algorithm for estimating the value function in entropy-regularized Markov decision processes and two-player games, given a generative model of the environment. SmoothCruiser makes use of the smoothness of the Bellman operator promoted by the regularization to achieve problem-independent sample complexity of order O(1/ϵ 4) for a desired accuracy ϵ, whereas for non-regularized settings there are no known algorithms with guaranteed polynomial sample complexity in the worst case.

7.1.1.1. Deep reinforcement learning

"I'm sorry Dave, I'm afraid I can't do that" Deep Q-Learning From Forbidden Actions, [42]

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

MERL: Multi-Head Reinforcement Learning, [39]

A common challenge in reinforcement learning is how to convert the agent's interactions with an environment into fast and robust learning. For instance, earlier work makes use of domain knowledge to improve existing reinforcement learning algorithms in complex tasks. While promising, previously acquired knowledge is often costly and challenging to scale up. Instead, we decide to consider problem knowledge with signals from quantities relevant to solve any task, e.g., self-performance assessment and accurate expectations. $\mathcal{V}^{ex}$ is such a quantity. It is the fraction of variance explained by the value function V and measures the discrepancy between V and the returns. Taking advantage of $\mathcal{V}^{ex}$, we propose MERL, a general framework for structuring reinforcement learning by injecting problem knowledge into policy gradient updates. As a result, the agent is not only optimized for a reward but learns using problem-focused quantities provided by MERL, applicable out-of-the-box to any task. In this paper: (a) We introduce and define MERL, the multi-head reinforcement learning framework we use throughout this work. (b) We conduct experiments across a variety of standard benchmark environments, including 9 continuous control tasks, where results show improved performance. (c) We demonstrate that MERL also improves transfer learning on a set of challenging pixel-based tasks. (d) We ponder how MERL tackles the problem of reward sparsity and better conditions the feature space of reinforcement learning agents.

Self-Educated Language Agent With Hindsight Experience Replay For Instruction Following, [45]

Language creates a compact representation of the world and allows the description of unlimited situations and objectives through compositionality. These properties make it a natural fit to guide the training of interactive agents as it could ease recurrent challenges in Reinforcement Learning such as sample complexity, generalization, or multi-tasking. Yet, it remains an open-problem to relate language and RL in even simple instruction following scenarios. Current methods rely on expert demonstrations, auxiliary losses, or inductive biases in neural architectures. In this paper, we propose an orthogonal approach called Textual Hindsight Experience Replay (THER) that extends the Hindsight Experience Replay approach to the language setting. Whenever the agent does not fulfill its instruction, THER learns to output a new directive that matches the agent trajectory, and it relabels the episode with a positive reward. To do so, THER learns to map a state into an instruction by using past successful trajectories, which removes the need to have external expert interventions to relabel episodes as in vanilla HER. We observe that this simple idea also initiates a learning synergy between language acquisition and policy learning on instruction following tasks in the BabyAI environment.

High-Dimensional Control Using Generalized Auxiliary Tasks, [47]

A long-standing challenge in reinforcement learning is the design of function approximations and efficient learning algorithms that provide agents with fast training, robust learning, and high performance in complex environments. To this end, the use of prior knowledge, while promising, is often costly and, in essence, challenging to scale up. In contrast, we consider problem knowledge signals, that are any relevant indicator useful to solve a task, e.g., metrics of uncertainty or proactive prediction of future states. Our framework consists of predicting such complementary quantities associated with self-performance assessment and accurate expectations. Therefore, policy and value functions are no longer only optimized for a reward but are learned using environment-agnostic quantities. We propose a generally applicable framework for structuring reinforcement learning by injecting problem knowledge in policy gradient updates. In this paper: (a) We introduce MERL, our multi-head reinforcement learning framework for generalized auxiliary tasks. (b) We conduct experiments across a variety of standard benchmark environments. Our results show that MERL improves performance for on- and off-policy methods. (c) We show that MERL also improves transfer learning on a set of challenging tasks. (d) We investigate how our approach addresses the problem of reward sparsity and pushes the function approximations into a better-constrained parameter configuration.

7.1.2. Multi-armed Bandit Theory

Asymptotically Optimal Algorithms for Budgeted Multiple Play Bandits, [15]

We study a generalization of the multi-armed bandit problem with multiple plays where there is a cost associated with pulling each arm and the agent has a budget at each time that dictates how much she can expect to spend. We derive an asymptotic regret lower bound for any uniformly efficient algorithm in our setting. We then study a variant of Thompson sampling for Bernoulli rewards and a variant of KL-UCB for both single-parameter exponential families and bounded, finitely supported rewards. We show these algorithms are asymptotically optimal, both in rateand leading problem-dependent constants, including in the thick margin setting where multiple arms fall on the decision boundary.

Non-Asymptotic Pure Exploration by Solving Games, [46]

Pure exploration (aka active testing) is the fundamental task of sequentially gathering information to answer a query about a stochastic environment. Good algorithms make few mistakes and take few samples. Lower bounds (for multi-armed bandit models with arms in an exponential family) reveal that the sample complexity is determined by the solution to an optimisation problem. The existing state of the art algorithms achieve asymptotic optimality by solving a plug-in estimate of that optimisation problem at each step. We interpret the optimisation problem as an unknown game, and propose sampling rules based on iterative strategies to estimate and converge to its saddle point. We apply no-regret learners to obtain the first finite confidence guarantees that are adapted to the exponential family and which apply to any pure exploration query and bandit structure. Moreover, our algorithms only use a best response oracle instead of fully solving the optimisation problem.

Rotting bandits are not harder than stochastic ones, [32]

In bandits, arms' distributions are stationary. This is often violated in practice, where rewards change over time. In applications as recommendation systems, online advertising, and crowdsourcing, the changes may be triggered by the pulls, so that the arms' rewards change as a function of the number of pulls. In this paper, we consider the specific case of non-parametric rotting bandits, where the expected reward of an arm may decrease every time it is pulled. We introduce the filtering on expanding window average (FEWA) algorithm that at each round constructs moving averages of increasing windows to identify arms that are more likely to return high rewards when pulled once more. We prove that, without any knowledge on the decreasing behavior of the arms, FEWA achieves similar anytime problem-dependent, O(log(KT)), and problem-independent, O($sqrtKT$), regret bounds of near-optimal stochastic algorithms as UCB1 of Auer et al. (2002a). This result substantially improves the prior result of Levine et al. (2017) which needed knowledge of the horizon and decaying parameters to achieve problem-independent bound of only O(K1/3T2/3). Finally, we report simulations confirming the theoretical improvements of FEWA.

7.1.3. Black-box Optimization

General parallel optimization without a metric, [34]

Hierarchical bandits are an approach for global optimization of extremely irregular functions. This paper provides new elements regarding POO, an adaptive meta-algorithm that does not require the knowledge of local smoothness of the target function. We first highlight the fact that the subroutine algorithm used in POO should have a small regret under the assumption of local smoothness with respect to the chosen partitioning, which is unknown if it is satisfied by the standard subroutine HOO. In this work, we establish such regret guarantee for HCT, which is another hierarchical optimistic optimization algorithm that needs to know the smoothness. This confirms the validity of POO. We show that POO can be used with HCT as a subroutine with a regret upper bound that matches the one of best-known algorithms using the knowledge of smoothness up to a $\sqrt{\log n}$ factor. On top of that, we propose a general wrapper, called GPO, that can cope with algorithms that only have simple regret guarantees. Finally, we complement our findings with experiments on difficult functions.

A simple dynamic bandit algorithm for hyper-parameter tuning, [33]

Hyper-parameter tuning is a major part of modern machine learning systems. The tuning itself can be seen as a sequential resource allocation problem. As such, methods for multi-armed bandits have been already applied. In this paper, we view hyper-parameter optimization as an instance of best-arm identification in infinitely many-armed bandits. We propose D-TTTS, a new adaptive algorithm inspired by Thompson sampling, which dynamically balances between refining the estimate of the quality of hyper-parameter configurations previously explored and adding new hyper-parameter configurations to the pool of candidates. The algorithm is easy to implement and shows competitive performance compared to state-of-the-art algorithms for hyper-parameter tuning.

7.1.4. Statistics for Machine Learning

Non-asymptotic analysis of a sequential rupture detection test and its application to non-stationary bandits, [36]

We study a strategy for online change-point detection based on generalized likelihood ratios (GLR) and that can be expressed with the binary relative entropy. This test is used to detect a change in the mean of a bounded distribution, and we propose a non-asymptotic control of its false alarm probability and detection delay. We then explain how it can be useful for sequential decision making by proposing the GLR-klUCB bandit strategy, which is efficient in piece-wise stationary multi-armed bandit models.

Sequential change-point detection: Laplace concentration of scan statistics and non-asymptotic delay bounds, [27]

We consider change-point detection in a fully sequential setup, when observations are received one by one and one must raise an alarm as early as possible after any change. We assume that both the change points and the distributions before and after the change are unknown. We consider the class of piecewise-constant mean processes with sub-Gaussian noise, and we target a detection strategy that is uniformly good on this class (this constrains the false alarm rate and detection delay). We introduce a novel tuning of the GLR test that takes here a simple form involving scan statistics, based on a novel sharp concentration inequality using an extension of the Laplace method for scan-statistics that holds doubly-uniformly in time. This also considerably simplifies the implementation of the test and analysis. We provide (perhaps surprisingly) the first fully non-asymptotic analysis of the detection delay of this test that matches the known existing asymptotic orders, with fully explicit numerical constants. Then, we extend this analysis to allow some changes that are not-detectable by any uniformly-good strategy (the number of observations before and after the change are too small for it to be detected by any such algorithm), and provide the first robust, finite-time analysis of the detection delay.

Learning Multiple Markov Chains via Adaptive Allocation, [35]

We study the problem of learning the transition matrices of a set of Markov chains from a single stream of observations on each chain. We assume that the Markov chains are ergodic but otherwise unknown. The learner

can sample Markov chains sequentially to observe their states. The goal of the learner is to sequentially select various chains to learn transition matrices uniformly well with respect to some loss function. We introduce a notion of loss that naturally extends the squared loss for learning distributions to the case of Markov chains, and further characterize the notion of being uniformly good in all problem instances. We present a novel learning algorithm that efficiently balances exploration and exploitation intrinsic to this problem, without any prior knowledge of the chains. We provide finite-sample PAC-type guarantees on the performance of the algorithm. Further, we show that our algorithm asymptotically attains an optimal loss.

7.1.5. DPP

On two ways to use determinantal point processes for Monte Carlo integration, [40]

This paper focuses on Monte Carlo integration with determinantal point processes (DPPs) which enforce negative dependence between quadrature nodes. We survey the properties of two unbiased Monte Carlo estimators of the integral of interest: a direct one proposed by Bardenet & Hardy (2016) and a less obvious 60-year-old estimator by Ermakov & Zolotukhin (1960) that actually also relies on DPPs. We provide an efficient implementation to sample exactly a particular multidimen-sional DPP called multivariate Jacobi ensemble. This let us investigate the behavior of both estima-tors on toy problems in yet unexplored regimes.

7.2. Applications

7.2.1. Autonomous car

Practical Open-Loop Optimistic Planning, [25]

We consider the problem of online planning in a Markov Decision Process when given only access to a generative model, restricted to open-loop policies-i.e. sequences of actions-and under budget constraint. In this setting, the Open-Loop Optimistic Planning (OLOP) algorithm enjoys good theoretical guarantees but is overly conservative in practice, as we show in numerical experiments. We propose a modified version of the algorithm with tighter upper-confidence bounds, KL-OLOP, that leads to better practical performances while retaining the sample complexity bound. Finally, we propose an efficient implementation that significantly improves the time complexity of both algorithms.

Budgeted Reinforcement Learning in Continuous State Space, [20]

A Budgeted Markov Decision Process (BMDP) is an extension of a Markov Decision Process to critical applications requiring safety constraints. It relies on a notion of risk implemented in the shape of a cost signal constrained to lie below an-adjustable-threshold. So far, BMDPs could only be solved in the case of finite state spaces with known dynamics. This work extends the state-of-the-art to continuous spaces environments and unknown dynamics. We show that the solution to a BMDP is a fixed point of a novel Budgeted Bellman Optimality operator. This observation allows us to introduce natural extensions of Deep Reinforcement Learning algorithms to address large-scale BMDPs. We validate our approach on two simulated applications: spoken dialogue and autonomous driving.

7.2.2. Cognitive radio

Decentralized Spectrum Learning for IoT Wireless Networks Collision Mitigation, [28]

This paper describes the principles and implementation results of reinforcement learning algorithms on IoT devices for radio collision mitigation in ISM unlicensed bands. Learning is here used to improve both the IoT network capability to support a larger number of objects as well as the autonomy of IoT devices. We first illustrate the efficiency of the proposed approach in a proof-of-concept based on USRP software radio platforms operating on real radio signals. It shows how collisions with other RF signals present in the ISM band are diminished for a given IoT device. Then we describe the first implementation of learning algorithms on LoRa devices operating in a real LoRaWAN network, that we named IoTligent. The proposed solution adds neither processing overhead so that it can be ran in the IoT devices, nor network overhead so that no change is required to LoRaWAN. Real life experiments have been done in a realistic LoRa network and they show that IoTligent device battery life can be extended by a factor 2 in the scenarios we faced during our experiment.

GNU Radio Implementation of MALIN: "Multi-Armed bandits Learning for Internet-of-things Networks", [37]

We implement an IoT network in the following way: one gateway, one or several intelligent (i.e., learning) objects, embedding the proposed solution, and a traffic generator that emulates radio interferences from many other objects. Intelligent objects communicate with the gateway with a wireless ALOHA-based protocol, which does not require any specific overhead for the learning. We model the network access as a discrete sequential decision making problem, and using the framework and algorithms from Multi-Armed Bandit (MAB) learning, we show that intelligent objects can improve their access to the network by using low complexity and decentralized algorithms, such as UCB1 and Thompson Sampling. This solution could be added in a straightforward and costless manner in LoRaWAN networks, just by adding this feature in some or all the devices, without any modification on the network side.

7.2.3. Other

Accurate reconstruction of EBSD datasets by a multimodal data approach using an evolutionary algorithm, [14]

A new method has been developed for the correction of the distortions and/or enhanced phase differentiation in Electron Backscatter Diffraction (EBSD) data. Using a multi-modal data approach, the method uses segmented images of the phase of interest (laths, precipitates, voids, inclusions) on images gathered by backscattered or secondary electrons of the same area as the EBSD map. The proposed approach then search for the best transformation to correct their relative distortions and recombines the data in a new EBSD file. Speckles of the features of interest are first segmented in both the EBSD and image data modes. The speckle extracted from the EBSD data is then meshed, and the Covariance Matrix Adaptation Evolution Strategy (CMA-ES) is implemented to distort the mesh until the speckles superimpose. The quality of the matching is quantified via a score that is linked to the number of overlapping pixels in the speckles. The locations of the points of the distorted mesh are compared to those of the initial positions to create pairs of matching points that are used to calculate the polynomial function that describes the distortion the best. This function is then applied to un-distort the EBSD data, and the phase information is inferred using the data of the segmented speckle. Fast and versatile, this method does not require any human annotation and can be applied to large datasets and wide areas. Besides, this method requires very few assumptions concerning the shape of the distortion function. It can be used for the single compensation of the distortions or combined with the phase differentiation. The accuracy of this method is of the order of the pixel size. Some application examples in multiphase materials with feature sizes down to 1 μm are presented, including Ti-6Al-4V Titanium alloy, Rene 65 and additive manufactured Inconel 718 Nickel-base superalloys.

Energy Management for Microgrids: a Reinforcement Learning Approach, [41]

This paper presents a framework based on reinforcement learning for energy management and economic dispatch of an islanded microgrid without any forecasting module. The architecture of the algorithm is divided in two parts: a learning phase trained by a reinforcement learning (RL) algorithm on a small dataset and the testing phase based on a decision tree induced from the trained RL. An advantage of this approach is to create an autonomous agent, able to react in real-time, considering only the past. This framework was tested on real data acquired at Ecole Polytechnique in France over a long period of time, with a large diversity in the type of days considered. It showed near optimal, efficient and stable results in each situation.

SPIRALS Project-Team

7. New Results

7.1. Browser fingerprinting

We obtained new results on the concept of browser fingerprinting. This is a major technique of Internet security that is widely used for many purposes such as tracking activities, enhancing authentication, detecting bots, just to name a few. These results contribute to the enhancement of security for distributed software systems.

Our contributions to browser fingerprinting include the following three elements. First, we collected 122K fingerprints from 2 346 browsers and studied their stability over more than 2 years. We showed that, despite frequent changes in the fingerprints, a significant fraction of browsers can be tracked over a long period of time. Second, we designed a test suite to evaluate fingerprinting countermeasures. We applied our test suite to 7 countermeasures, some of them claiming to generate consistent fingerprints, and show that all of them can be identified, which can make their users more identifiable. Third, we explored the use of browser fingerprinting for crawler detection. We measured its use in the wild, as well as the main detection techniques. Since fingerprints are collected on the client-side, we also evaluated its resilience against an adversarial crawler developer that tries to modify its crawler fingerprints to bypass security checks.

These results have been obtained in the context of the PhD thesis of Antoine Vastel [14] defended in October 2019.

7.2. Test amplification

With respect to self-healing, we proposed a new algorithm for test amplification. Test amplification consists of exploiting the knowledge of test methods, in which developers embed input data and expected properties, in order to enhance these tests [22].

We proposed a new approach based on test inputs transformation and assertions generation to amplify test suites, and implemented this approach in the DSpot software tool that we created [21]. By evaluating DSpot on open-source projects from GitHub, we showed that we improve the mutation score of test suites. These improvements have been proposed to developers through pull requests: their feedbacks show that they value the output of DSpot by accepting to integrate amplified test methods into their test suite. This proves that DSpot can improve the quality of the test suite of real projects. We also showed that DSpot can generate amplified test methods that specify behavioral changes, and can generate amplified test methods to improve the ability to detect potential regressions.

These results have been obtained in the context of the STAMP H2020 project and in the context of the PhD thesis of Benjamin Danglot [11] defended in November 2019.

7.3. Understanding mobile-specific code smells

With respect to self-healing, we obtained new results in the domain of code smells for mobile software systems. Code smells are well-known concepts in software engineering. They refer to bad design and development practices commonly observed in software systems.

We obtained three new results that contribute to a better understanding of mobile code smells. First, we studied the expansion of code smells in different mobile platforms. Then, we conducted a large-scale study to analyze the change history of mobile apps and discern the factors that favor the introduction and survival of code smells. To consolidate these studies, we also performed a user study to investigate developers' perception of code smells and the adequacy of static analyzers as a solution for coping with them. Finally, we performed a qualitative study to question the established foundation about the definition and detection of mobile code smells. The results of these studies revealed important research findings. Notably, we showed that pragmatism, prioritization, and individual attitudes are not relevant factors for the accrual of mobile code smells. The problem is rather caused by ignorance and oversight, which are prevalent among mobile developers. Furthermore, we highlighted several flaws in the code smell definitions that are currently adopted by the research community. These results allowed us to elaborate some recommendations for researchers and tool makers willing to design detection and refactoring tools for mobile code smells [33], [34]. On top of that, our results opened perspectives for research works about the identification of mobile code smells and development practices in general.

These results have been obtained in the context of the PhD thesis of Sarra Habchi [12] defended in December 2019.

7.4. Towards privacy-sensitive mobile crowdsourcing

We obtained new results in the domain of data privacy for crowdsourced data.

We proposed an anonymous data collection library for mobile apps, a software library that improves the user's privacy without compromising the overall quality of the crowdsourced dataset. In particular, we proposed a decentralized approach, named FOUGERE, to convey data samples from user devices using peer-to-peer (P2P) communications to third-party servers, thus introducing an a priori data anonymization process that is resilient to location-based attacks. To validate the approach, we proposed a testing framework to test this P2P communication library, named PeerFleet. Beyond the identification of P2P-related errors, PeerFleet also helps to tune the discovery protocol settings to optimize the deployment of P2P apps. We validated FOUGERE using 500 emulated devices that replay a mobility dataset and use FOUGERE to collect location data. We evaluated the overhead, the privacy and the utility of FOUGERE. We showed that FOUGERE defeats the state-of-the-art location-based privacy attacks with little impact on the quality of the collected data [38], [5].

These results have been obtained in the context of the PhD thesis of Lakhdar Meftah [13] defended in December 2019.

VALSE Project-Team

7. New Results

7.1. Stabilization of multistable systems

Participants: Rosane Ushirobira, Denis Efimov.

The problem of robust stabilization of affine nonlinear multistable systems with respect to disturbance inputs was studied in [29]. The results are obtained using the framework of input-to-state stability and integral input-to-state stability for systems with multiple invariant sets. The notions of corresponding control Lyapunov functions as well as the small control property are extended within the multistability framework. It is verified that the universal control formula can be applied. In a similar vein, input-to-state stability and stabilization of passive multistable systems were investigated in [53].

7.2. Robustness of homogeneous systems

Participants: Andrey Polyakov, Denis Efimov.

In [32], we studied the finite-time stability of a class of nonlinear systems $\dot{x} = f(x) = H(x)b(x)$, where H is homogeneous and b is bounded. We defined the homogeneous extension of the non-homogeneous function f and used this extension to prove that, under some conditions on b, if the system $\dot{x} = f(x)$ is globally asymptotically stable, then it is finite-time stable. In [31], a theoretical basement of the previous result has been given showing robust stability of the system $\dot{x} = f(x) = H(x)b(x)$ by considering b as a perturbation.

7.3. Finite-time control and estimation in chemostat

Participants: Denis Efimov, Andrey Polyakov.

In [34], the problem of robust stabilization of the concentration of two different species competing for a single limiting substrate was addressed. Such a stabilization is performed by means of discontinuous feedback control laws that ensure coexistence of all species. The control laws are designed considering bounded uncertainties on the kinetic rates and full state measurements. The problem of robust finite-time estimation of the state for a chemostat was considered in [35].

7.4. LKF for neutral type time-delay systems

Participant: Denis Efimov.

The problem of existence of a Lyapunov-Krasovskii functional (LKF) for nonlinear neutral type time-delay systems was revisited in [37] considering the uniform stability analysis and the LKF in a Sobolev space of absolutely continuous functions with bounded derivatives.

7.5. Finite-time/Fixed-time control of PDEs

Participant: Andrey Polyakov.

The paper [40] deals with continuous boundary time-varying feedbacks for fixed-time stabilization of constant-parameter reaction-diffusion systems. IT was shown that the time of convergence can be prescribed and is independent of the initial condition of the system. The design of time-varying feedbacks was carried out by using the backstepping approach for which suitable characterizations for time-varying kernels were derived. The work [39] considered the problem of local finite-time stabilization of the viscous Burgers equation. A boundary switched linear control with state dependent switching law was designed based on the backstepping approach. The strategy builds on discontinuous kernels, which render the control function to be piecewise continuous. It was proven that such a control stabilizes locally the viscous Burgers equation and that the settling time depends on initial conditions.

7.6. Interval prediction for LPV systems

Participant: Denis Efimov.

The problem of behaviour prediction for linear parameter-varying (LPV) systems was considered in the interval framework in [41]. It was assumed that the system is subject to uncertain inputs and the vector of scheduling parameters is unmeasurable, but all uncertainties take values in a given admissible set. Then an interval predictor was designed and its stability was guaranteed applying Lyapunov function with a novel structure. The conditions of stability were formulated in the form of linear matrix inequalities. Efficiency of the theoretical results was demonstrated in the application to safe motion planning for autonomous vehicles (see videos of numeric experiments).

7.7. Finite-time/Fixed-time stability and stabilization

Participants: Andrey Polyakov, Rosane Ushirobira.

In [33], we studied the issues of prescribed-time stabilization of a chain of integrators of arbitrary length, that can be either pure (i.e. with no disturbance) or perturbed. The feedback law proposed by Song et al. was revisited, and it was shown that by an appropriately recast within the framework of time-varying homogeneity it can stabilize the system in a prescribed time. Characterizations (necessary and sufficient conditions) of finite-time and fixed-time stability of evolution inclusions in Banach spaces were presented in [44] in terms of Lyapunov functionals.

7.8. Robust stability analysis for Persidskii systems

Participant: Denis Efimov.

A class of generalized nonlinear Persidskii systems was considered in [36]. The conditions of input-to-state and integral input-to-state stability were established, which can be checked using linear matrix inequalities. The issues of discretization of this class of dynamics were analyzed using the Euler methods. The proposed theory was applied to a Lotka-Volterra model.

7.9. Design of a distributed finite-time observer

Participants: Denis Efimov, Rosane Ushirobira.

In [46], a distributed observer was presented to estimate the state of a linear time-invariant plant in finite-time in each observer node. The design was based on a decomposition into locally observable and unobservable substates and on properties of homogeneous systems. Each observer node can reconstruct in finite-time its locally observable substate with its measurements only. Then exploiting the coupling, a finite-time converging observer was constructed for the remaining states by adding the consensus terms.

7.10. Parameter estimation in finite-time without PE

Participant: Denis Efimov.

The problem of estimation in the linear regression model was studied in [24], [48], [25] under the hypothesis that the regressor may be excited on a limited initial interval of time only. Then the estimation solution was searched on a finite interval of time also based on the framework of finite-time or fixed-time converging dynamical systems. The robustness issue was analyzed and a short-time input-to-state stability property was introduced for fixed-time converging time-varying systems with a sufficient condition, which was formulated with the use of a Lyapunov function. Several estimation algorithms were proposed and compared with existing solutions.

7.11. Quadrotor control

Participant: Andrey Polyakov.

The problem of a state feedback design for control of a quadrotor system under state and time constraints was studied in [49]. Convex embedding approach and Implicit Lyapunov function were employed to design a finite-time controller. The feedback gains were calculated by a system of LMIs (see video of experiments).

7.12. Blimp robot control

Participant: Denis Efimov.

The papers [50], [26] presented the robust controller design for an indoor blimp robot to achieve application such as the surveillance. The commonly used 6 degrees of freedom dynamic model was simplified under reasonable assumptions and decoupled into two independent parts. The blimp simplified horizontal plane movement model was complemented with disturbance terms to ensure the modeling accuracy, then it was transformed to a simpler form for the ease of controller design. Next, the disturbance terms were evaluated by the designed real-time estimator, and the perturbation estimates were compensated in the conceived motion controller for cancellation of the influence of disturbances. The performance and robustness of the disturbance compensation-based controller were verified by both simulations and experiments on the developed blimp robot.

7.13. Output finite-time stability and stabilization

Participants: Denis Efimov, Andrey Polyakov.

The equivalent Lyapunov characterizations of output finite-time stability were presented in [51]. Another sufficient condition for output finite-time stability was presented in [52]. Based on this condition a scheme of adaptive finite-time control design was provided. The presented results were obtained with the use of homogeneity property.