RESEARCH CENTRE

Inria Paris Centre

IN PARTNERSHIP WITH:

Ecole normale supérieure de Paris, CNRS

2023
ACTIVITY REPORT

Project-Team
SIERRA

Statistical Machine Learning and Parsimony

IN COLLABORATION WITH: Département d'Informatique de l'Ecole Normale Supérieure

DOMAIN

Applied Mathematics, Computation and Simulation

THEME

Optimization, machine learning and statistical methods

Contents

Project-Team SIERRA

Creation of the Project-Team: 2012 January 01

Keywords

Computer sciences and digital sciences

A3.4. – Machine learning and statistics

A5.4. – Computer vision

A6.2. – Scientific computing, Numerical Analysis & Optimization

A7.1. – Algorithms

A8.2. – Optimization

A9.2. – Machine learning

Other research topics and application domains

B9.5.6. – Data science

1 Team members, visitors, external collaborators

Research Scientists

- Francis Bach [Team leader, INRIA, HDR]
- Michael Jordan [Fondation Inria, from Jun 2023]
- Alessandro Rudi [INRIA, Researcher]
- Umut Simsekli [INRIA, Researcher]
- Adrien Taylor [INRIA, Researcher, from Oct 2023]
- Alexandre d'Aspremont [CNRS, Senior Researcher, HDR]

Post-Doctoral Fellows

- Pierre-Cyril Aubin [INRIA, Post-Doctoral Fellow, until Aug 2023]
- Luc Brogat-Motte [CENTRALESUPELEC, Post-Doctoral Fellow, from Oct 2023]
- Fajwel Fogel [ENS PARIS, Post-Doctoral Fellow, from Sep 2023]
- David Holzmuller [INRIA, Post-Doctoral Fellow, from Nov 2023]
- Ziad Kobeissi [DI-ENS, Post-Doctoral Fellow, until Sep 2023]
- Clément Mantoux [INRIA, Post-Doctoral Fellow, until Aug 2023]
- Anant Raj [INRIA, Post-Doctoral Fellow]
- Corbinian Schlosser [INRIA, Post-Doctoral Fellow, from Jul 2023]
- Yang Su [CEA, from Sep 2023]
- Paul Viallard [INRIA, Post-Doctoral Fellow, from Feb 2023]
- Blake Woodworth [INRIA, Post-Doctoral Fellow, until Jun 2023]

PhD Students

- Antoine Bambade [Ecole des Ponts]
- Melih Barsbey [UNIV BOGAZICI, from Aug 2023]
- Andrea Basteri [INRIA]
- Gaspard Beugnot [INRIA]
- Pierre Boudart [INRIA, from Feb 2023]
- Sarah Brood [ENS Paris, from Nov 2023]
- Arthur Calvi [CNRS]
- Theophile Cantelobre [INRIA]
- Benjamin Dupuis [INRIA, from May 2023]
- Bertille Follain [ENS PARIS]
- Gautier Izacard [CNRS, until Jan 2023]
- Remi Jezequel [ENS Paris, until Jan 2023]

- Marc Lambert [DGA]
- Clément Lezane [UNIV TWENTE]
- Simon Martin [INRIA, from Sep 2023]
- Céline Moucer [ENS PARIS-SACLAY]
- Benjamin Paul-Dubois-Taine [UNIV PARIS SACLAY]
- Dario Shariatian [INRIA, from Oct 2023]
- Lawrence Stewart [INRIA]

Interns and Apprentices

- Eugene Berta [INRIA, Intern, from Apr 2023 until Sep 2023]
- Matthieu Dinot [INRIA, Intern, from Apr 2023 until Aug 2023]
- Benjamin Dupuis [INRIA, Intern, until Mar 2023]
- Krunoslav Lehman Pavasovic [INRIA, Intern, until Mar 2023]
- Simon Martin [ENS Paris, Intern, from Apr 2023 until Aug 2023]
- Sarah Sachs [INRIA, Intern, until Feb 2023]

Administrative Assistants

- Meriem Guemair [INRIA]
- Marina Kovacic [Inria, from Aug 2023]

Visiting Scientists

- Silvere Bonnabel [Ecole des Mines de Paris, from Aug 2023]
- Laurent El Ghaoui [UNIV BERKELEY, from Jun 2023 until Jun 2023, HDR]
- Steffen Grunewalder [UNIV NEWCASTLE, until Jun 2023]
- Cristobal Guzman [Catholic University of Chile]
- Max Kramkimel [NIC, from Mar 2023 until Jul 2023]
- Antônio Horta Ribeiro [UNIV UPPSALA, from Mar 2023 until Jun 2023]

2 Overall objectives

2.1 Statement

Machine learning is a recent scientific domain, positioned between applied mathematics, statistics and computer science. Its goals are the optimization, control, and modelisation of complex systems from examples. It applies to data from numerous engineering and scientific fields (e.g., vision, bioinformatics, neuroscience, audio processing, text processing, economy, finance, etc.), the ultimate goal being to derive general theories and algorithms allowing advances in each of these domains. Machine learning is characterized by the high quality and quantity of the exchanges between theory, algorithms and applications: interesting theoretical problems almost always emerge from applications, while theoretical analysis allows the understanding of why and when popular or successful algorithms do or do not work, and leads to proposing significant improvements.

Our academic positioning is exactly at the intersection between these three aspects—algorithms, theory and applications—and our main research goal is to make the link between theory and algorithms, and between algorithms and high-impact applications in various engineering and scientific fields, in particular computer vision, bioinformatics, audio processing, and text processing.

3 Research program

Machine learning has emerged as its own scientific domain in the last 30 years, providing a good abstraction of many problems and allowing exchanges of best practices between data oriented scientific fields. Among its main research areas, there are currently probabilistic models, supervised learning (including neural networks), unsupervised learning, reinforcement learning, and statistical learning theory. All of these are represented in the SIERRA team, but the main goals of the team are mostly related to supervised learning and optimization, and their mutual interactions, as well as with interdisciplinary collaborations. One particularity of the team is the strong focus on optimization (in particular convex optimization, but with more works in the non-convex world recently), leading to contributions in optimization which go beyond the machine learning context.

We have thus divided our research effort in three parts:

1. Convex optimization
2. Non-convex optimization
3. Machine learning.

4 Application domains

Machine learning research can be conducted from two main perspectives: the first one, which has been dominant in the last 30 years, is to design learning algorithms and theories which are as generic as possible, the goal being to make as few assumptions as possible regarding the problems to be solved and to let data speak for themselves. This has led to many interesting methodological developments and successful applications. However, we believe that this strategy has reached its limit for many application domains, such as computer vision, bioinformatics, neuro-imaging, text and audio processing, which leads to the second perspective our team is built on: Research in machine learning theory and algorithms should be driven by interdisciplinary collaborations, so that specific prior knowledge may be properly introduced into the learning process, in particular with the following fields:

- Computer vision: object recognition, object detection, image segmentation, image/video processing, computational photography. In collaboration with the Willow project-team.
- Bioinformatics: cancer diagnosis, protein function prediction, virtual screening.
- Text processing: document collection modeling, language models.
- Audio processing: source separation, speech/music processing.
- Climate science (satellite imaging).

5 Social and environmental responsibility

As one domain within applied mathematics and computer science, machine learning and artificial intelligence may contribute positively to the environment for example by measuring climate change effect or reducing the carbon footprint of other sciences and activities. But it may also contribute negatively, notably by the ever-increasing sizes of machine learning models. Within the team, we work on these two aspects through our work on climate science and on frugal algorithms.

6 Highlights of the year

Recruiting. We happy to welcome to new permanent researches, Michael Jordan and Adrien Taylor.

Team renewal. Our team has been evaluted this year, both by Inria and HCERES (through our ENS affiliation). Since we reach the end of the 12 year cycle, we have proposed a re-creation of the team, which is currently being processed.

7 New results

7.1 Solving moment and polynomial optimization problems on Sobolev spaces

Using standard tools of harmonic analysis, we state and solve the problem of moments for positive measures supported on the unit ball of a Sobolev space of multivariate periodic trigonometric functions. We describe outer and inner semidefinite approximations of the cone of Sobolev moments. They are the basic components of an infinite-dimensional moment-sums of squares hierarchy, allowing to solve numerically non-convex polynomial optimization problems on infinite-dimensional Sobolev spaces, with global convergence guarantees.

7.2 GloptiNets: Scalable Non-Convex Optimization with Certificates

We present a novel approach to non-convex optimization with certificates, which handles smooth functions on the hypercube or on the torus. Unlike traditional methods that rely on algebraic properties, our algorithm exploits the regularity of the target function intrinsic in the decay of its Fourier spectrum. In [28] we show that by defining a tractable family of models, we allow at the same time to obtain precise certificates and to leverage the advanced and powerful computational techniques developed to optimize neural networks. In this way the scalability of our approach is naturally enhanced by parallel computing with GPUs. Our approach, when applied to the case of polynomials of moderate dimensions but with thousands of coefficients, outperforms the state-of-the-art optimization methods with certificates, as the ones based on Lasserre's hierarchy, addressing problems intractable for the competitors.

7.3 Non-Parametric Learning of Stochastic Differential Equations with Fast Rates of Convergence

In [44] we propose a novel non-parametric learning paradigm for the identification of drift and diffusion coefficients of non-linear stochastic differential equations, which relies upon discrete-time observations of the state. The key idea essentially consists of fitting a RKHS-based approximation of the corresponding Fokker-Planck equation to such observations, yielding theoretical estimates of learning rates which, unlike previous works, become increasingly tighter when the regularity of the unknown drift and diffusion coefficients becomes higher. Our method being kernel-based, offline pre-processing may in principle be profitably leveraged to enable efficient numerical implementation.

7.4 Efficient Sampling of Stochastic Differential Equations with Positive Semi-Definite Models

This line of works deals with the problem of efficient sampling from a stochastic differential equation, given the drift function and the diffusion matrix. The proposed approach leverages a recent model for probabilities (the positive semi-definite – PSD model) from which it is possible to obtain independent and identically distributed (i.i.d.) samples at precision ϵ with a cost that is $m^2 d \log(1/\epsilon)$ where m is the dimension of the model, d the dimension of the space. The proposed approach consists in: first, computing the PSD model that satisfies the Fokker-Planck equation (or its fractional variant) associated with the SDE, up to error ϵ, and then sampling from the resulting PSD model. Assuming some regularity of the Fokker-Planck solution (i.e. β-times differentiability plus some geometric condition on its zeros) We obtain an algorithm that: (a) in the preparatory phase obtains a PSD model with L2 distance ϵ from the solution of the equation, with a model of dimension $m = \epsilon^{-(d+1)/(\beta-2s)}(\log(1/\epsilon))^{d+1}$ where $1/2 \leq s \leq 1$ is the fractional power to the Laplacian, and total computational complexity of $O(m^{3.5}\log(1/\epsilon))$ and then (b) for Fokker-Planck equation, it is able to produce i.i.d. samples with error ϵ in Wasserstein-1 distance, with a cost that is $O(d\epsilon^{-2(d+1)/\beta-2}\log(1/\epsilon)^{2d+3})$ per sample. This means that, if the probability associated with the SDE is somewhat regular, i.e. $\beta \geq 4d+2$, then the algorithm requires $O(\epsilon^{-0.88}\log(1/\epsilon)^{4.5d})$ in the preparatory phase, and $O(\epsilon^{-1/2}\log(1/\epsilon)^{2d+2})$ for each sample. Our results suggest that as the true solution gets smoother, we can circumvent the curse of dimensionality without requiring any sort of convexity.

7.5 Automated tight Lyapunov analysis for first-order methods

We present a methodology for establishing the existence of quadratic Lyapunov inequalities for a wide range of first-order methods used to solve convex optimization problems. In particular, we consider i) classes of optimization problems of finite-sum form with (possibly strongly) convex and possibly smooth functional components, ii) first-order methods that can be written as a linear system on state-space form in feedback interconnection with the subdifferentials of the functional components of the objective function, and iii) quadratic Lyapunov inequalities that can be used to draw convergence conclusions. We provide a necessary and sufficient condition for the existence of a quadratic Lyapunov inequality that amounts to solving a small-sized semidefinite program. We showcase our methodology on several first-order methods that fit the framework. Most notably, our methodology allows us to significantly extend the region of parameter choices that allow for duality gap convergence in the Chambolle-Pock method when the linear operator is the identity mapping.

7.6 Provable non-accelerations of the heavy-ball method

In this work, we show that the heavy-ball (HB) method provably does not reach an accelerated convergence rate on smooth strongly convex problems. More specifically, we show that for any condition number and any choice of algorithmic parameters, either the worst-case convergence rate of HB on the class of -smooth and -strongly convex *quadratic* functions is not accelerated (that is, slower than), or there exists an -smooth -strongly convex function and an initialization such that the method does not converge. To the best of our knowledge, this result closes a simple yet open question on one of the most used and iconic first-order optimization technique. Our approach builds on finding functions for which HB fails to converge and instead cycles over finitely many iterates. We analytically describe all parametrizations of HB that exhibit this cycling behavior on a particular cycle shape, whose choice is supported by a systematic and constructive approach to the study of cycling behaviors of first-order methods. We show the robustness of our results to perturbations of the cycle, and extend them to class of functions that also satisfy higher-order regularity conditions.

7.7 Sum-of-Squares Relaxations for Information Theory and Variational Inference

We consider extensions of the Shannon relative entropy, referred to as f-divergences. Three classical related computational problems are typically associated with these divergences: (a) estimation from moments, (b) computing normalizing integrals, and (c) variational inference in probabilistic models. These problems are related to one another through convex duality, and for all them, there are many applications throughout data science, and we aim for computationally tractable approximation algorithms that preserve properties of the original problem such as potential convexity or monotonicity. In order to achieve this, we derive a sequence of convex relaxations for computing these divergences from non-centered covariance matrices associated with a given feature vector: starting from the typically non-tractable optimal lower-bound, we consider an additional relaxation based on “sums-of-squares”, which is is now computable in polynomial time as a semidefinite program. We also provide computationally more efficient relaxations based on spectral information divergences from quantum information theory. For all of the tasks above, beyond proposing new relaxations, we derive tractable convex optimization algorithms, and we present illustrations on multivariate trigonometric polynomials and functions on the Boolean hypercube

7.8 Two losses are better than one: Faster optimization using a cheaper proxy

We present an algorithm for minimizing an objective with hard-to-compute gradients by using a related, easier-to-access function as a proxy. Our algorithm is based on approximate proximalpoint iterations on the proxy combined with relatively few stochastic gradients from the objective. When the difference between the objective and the proxy is δ-smooth, our algorithm guarantees convergence at a rate matching stochastic gradient descent on a δ-smooth objective, which can lead to substantially better sample efficiency. Our algorithm has many potential applications in machine learning, and provides a principled means of leveraging synthetic data, physics simulators, mixed public and private data, and more.

7.9 Classifier Calibration with ROC-Regularized Isotonic Regression

Calibration of machine learning classifiers is necessary to obtain reliable and interpretable predictions, bridging the gap between model confidence and actual probabilities. One prominent technique, isotonic regression (IR), aims at calibrating binary classifiers by minimizing the cross entropy on a calibration set via monotone transformations. IR acts as an adaptive binning procedure, which allows achieving a calibration error of zero, but leaves open the issue of the effect on performance. In this line of work, we first prove that IR preserves the convex hull of the ROC curve—an essential performance metric for binary classifiers. This ensures that a classifier is calibrated while controlling for overfitting of the calibration set. We then present a novel generalization of isotonic regression to accommodate classifiers with K classes. Our method constructs a multidimensional adaptive binning scheme on the probability simplex, again achieving a multi-class calibration error equal to zero. We regularize this algorithm by imposing a form of monotony that preserves the K-dimensional ROC surface of the classifier. We show empirically that this general monotony criterion is effective in striking a balance between reducing cross entropy loss and avoiding overfitting of the calibration set.

7.10 Regularization properties of adversarially-trained linear regression

State-of-the-art machine learning models can be vulnerable to very small input perturbations that are adversarially constructed. Adversarial training is an effective approach to defend against it. Formulated as a min-max problem, it searches for the best solution when the training data were corrupted by the worst-case attacks. Linear models are among the simple models where vulnerabilities can be observed and are the focus of our study. In this case, adversarial training leads to a convex optimization problem which can be formulated as the minimization of a finite sum. We provide a comparative analysis between the solution of adversarial training in linear regression and other regularization methods. Our main findings are that: (A) Adversarial training yields the minimum-norm interpolating solution in the overparameterized regime (more parameters than data), as long as the maximum disturbance radius is smaller than a threshold. And, conversely, the minimumnorm interpolator is the solution to adversarial training with a given radius. (B) Adversarial training can be equivalent to parameter shrinking methods (ridge regression and Lasso). This happens in the underparametrized region, for an appropriate choice of adversarial radius and zero-mean symmetrically distributed covariates. (C) For ℓ_∞-adversarial training—as in square-root Lasso—the choice of adversarial radius for optimal bounds does not depend on the additive noise variance. We confirm our theoretical findings with numerical examples.

7.11 Differentiable Clustering with Perturbed Spanning Forests

We introduce a differentiable clustering method based on stochastic perturbations of minimum-weight spanning forests. This allows us to include clustering in end-toend trainable pipelines, with efficient gradients. We show that our method performs well even in difficult settings, such as data sets with high noise and challenging geometries. We also formulate an ad hoc loss to efficiently learn from partial clustering data using this operation. We demonstrate its performance on several data sets for supervised and semi-supervised tasks.

7.12 On the impact of activation and normalization in obtaining isometric embeddings at initialization

In this work, we explore the structure of the penultimate Gram matrix in deep neural networks, which contains the pairwise inner products of outputs corresponding to a batch of inputs. In several architectures it has been observed that this Gram matrix becomes degenerate with depth at initialization, which dramatically slows training. Normalization layers, such as batch or layer normalization, play a pivotal role in preventing the rank collapse issue. Despite promising advances, the existing theoretical results do not extend to layer normalization, which is widely used in transformers, and can not quantitatively characterize the role of non-linear activations. To bridge this gap, we prove that layer normalization, in conjunction with activation layers, biases the Gram matrix of a multilayer perceptron towards the identity matrix at an exponential rate with depth at initialization. We quantify this rate using the Hermite expansion of the activation function.

7.13 Kernelized diffusion maps

Spectral clustering and diffusion maps are celebrated dimensionality reduction algorithms built on eigenelements related to the diffusive structure of the data. The core of these procedures is the approximation of a Laplacian through a graph kernel approach, however this local average construction is known to be cursed by the high-dimension d. In this paper, we build a different estimator of the Laplacian's eigenvectors, via a reproducing kernel Hilbert space method, which adapts naturally to the regularity of the problem. We provide non-asymptotic statistical rates proving that the kernel estimator we build can circumvent the curse of dimensionality when the problem is well conditioned. Finally we discuss techniques (Nystrom subsampling, Fourier features) that enable to reduce the computational cost ¨ of the estimator while not degrading its overall performance

7.14 Convergence rates for non-log-concave sampling and log-partition estimation

Sampling from Gibbs distributions $p(x) \propto \exp(-V(x)/\varepsilon)$ and computing their log-partition function are fundamental tasks in statistics, machine learning, and statistical physics. However, while efficient algorithms are known for convex potentials V , the situation is much more difficult in the non-convex case, where algorithms necessarily suffer from the curse of dimensionality in the worst case. For optimization, which can be seen as a low-temperature limit of sampling, it is known that smooth functions V allow faster convergence rates. Specifically, for m-times differentiable functions in d dimensions, the optimal rate for algorithms with n function evaluations is known to be $O(n^{-m/d}$, where the constant can potentially depend on m, d and the function to be optimized. Hence, the curse of dimensionality can be alleviated for smooth functions at least in terms of the convergence rate. Recently, it has been shown that similarly fast rates can also be achieved with polynomial runtime $O(n^{3.5})$, where the exponent 3.5 is independent of m or d. Hence, it is natural to ask whether similar rates for sampling and log-partition computation are possible, and whether they can be realized in polynomial time with an exponent independent of m and d. We show that the optimal rates for sampling and log-partition computation are sometimes equal and sometimes faster than for optimization. We then analyze various polynomial-time sampling algorithms, including an extension of a recent promising optimization approach, and find that they sometimes exhibit interesting behavior but no near-optimal rates. Our results also give further insights on the relation between sampling, log-partition, and optimization problems.

7.15 Nonparametric Linear Feature Learning in Regression Through Regularisation

Representation learning plays a crucial role in automated feature selection, particularly in the context of high-dimensional data, where non-parametric methods often struggle. In this study, we focus on supervised learning scenarios where the pertinent information resides within a lower-dimensional linear subspace of the data, namely the multi-index model. If this subspace were known, it would greatly enhance prediction, computation, and interpretation. To address this challenge, we propose a novel method for linear feature learning with non-parametric prediction, which simultaneously estimates the prediction function and the linear subspace. Our approach employs empirical risk minimisation, augmented with a penalty on function derivatives, ensuring versatility. Leveraging the orthogonality and rotation invariance properties of Hermite polynomials, we introduce our estimator, named RegFeaL. By utilising alternative minimisation, we iteratively rotate the data to improve alignment with leading directions and accurately estimate the relevant dimension in practical settings. We establish that our method yields a consistent estimator of the prediction function with explicit rates. Additionally, we provide empirical results demonstrating the performance of RegFeaL in various experiments.

7.16 Approximate Heavy Tails in Offline (Multi-Pass) Stochastic Gradient Descent

A recent line of empirical studies has demonstrated that SGD might exhibit a heavy-tailed behavior in practical settings, and the heaviness of the tails might correlate with the overall performance. In this work, we investigate the emergence of such heavy tails. Previous works on this problem only considered, up to our knowledge, online (also called single-pass) SGD, in which the emergence of heavy tails in theoretical findings is contingent upon access to an infinite amount of data. Hence, the underlying mechanism generating the reported heavy-tailed behavior in practical settings, where the amount of training data is finite, is still not well-understood. Our contribution aims to fill this gap. In particular, we show that the stationary distribution of offline (also called multi-pass) SGD exhibits 'approximate' power-law tails and the approximation error is controlled by how fast the empirical distribution of the training data converges to the true underlying data distribution in the Wasserstein metric. Our main takeaway is that, as the number of data points increases, offline SGD will behave increasingly 'power-law-like'. To achieve this result, we first prove nonasymptotic Wasserstein convergence bounds for offline SGD to online SGD as the number of data points increases, which can be interesting on their own. Finally, we illustrate our theory on various experiments conducted on synthetic data and neural networks. Further details are in [12].

7.17 Uniform-in-Time Wasserstein Stability Bounds for (Noisy) Stochastic Gradient Descent

Algorithmic stability is an important notion that has proven powerful for deriving generalization bounds for practical algorithms. The last decade has witnessed an increasing number of stability bounds for different algorithms applied on different classes of loss functions. While these bounds have illuminated various properties of optimization algorithms, the analysis of each case typically required a different proof technique with significantly different mathematical tools. In this study, we make a novel connection between learning theory and applied probability and introduce a unified guideline for proving Wasserstein stability bounds for stochastic optimization algorithms. We illustrate our approach on stochastic gradient descent (SGD) and we obtain time-uniform stability bounds (i.e., the bound does not increase with the number of iterations) for strongly convex losses and nonconvex losses with additive noise, where we recover similar results to the prior art or extend them to more general cases by using a single proof technique. Our approach is flexible and can be generalizable to other popular optimizers, as it mainly requires developing Lyapunov functions, which are often readily available in the literature. It also illustrates that ergodicity is an important component for obtaining time-uniform bounds – which might not be achieved for convex or non-convex losses unless additional noise is injected to the iterates. Finally, we slightly stretch our analysis technique and prove time-uniform bounds for SGD under convex and non-convex losses (without additional additive noise), which, to our knowledge, is novel. Further information is in [21].

7.18 Learning via Wasserstein-Based High Probability Generalisation Bounds

Minimising upper bounds on the population risk or the generalisation gap has been widely used in structural risk minimisation (SRM) – this is in particular at the core of PAC-Bayesian learning. Despite its successes and unfailing surge of interest in recent years, a limitation of the PAC-Bayesian framework is that most bounds involve a Kullback-Leibler (KL) divergence term (or its variations), which might exhibit erratic behavior and fail to capture the underlying geometric structure of the learning problem – hence restricting its use in practical applications. As a remedy, recent studies have attempted to replace the KL divergence in the PAC-Bayesian bounds with the Wasserstein distance. Even though these bounds alleviated the aforementioned issues to a certain extent, they either hold in expectation, are for bounded losses, or are nontrivial to minimize in an SRM framework. In this work, we contribute to this line of research and prove novel Wasserstein distance-based PAC-Bayesian generalisation bounds for both batch learning with independent and identically distributed (i.i.d.) data, and online learning with potentially non-i.i.d. data. Contrary to previous art, our bounds are stronger in the sense that (i) they hold with high probability, (ii) they apply to unbounded (potentially heavy-tailed) losses, and (iii) they lead to optimizable training objectives that can be used in SRM. As a result we derive novel Wasserstein-based PAC-Bayesian learning algorithms and we illustrate their empirical advantage on a variety of experiments. More information can be found in [19].

7.19 Efficient Sampling of Stochastic Differential Equations with Positive Semi-Definite Models

This work deals with the problem of efficient sampling from a stochastic differential equation, given the drift function and the diffusion matrix. The proposed approach leverages a recent model for probabilities (the positive semi-definite – PSD model) from which it is possible to obtain independent and identically distributed (i.i.d.) samples at precision ϵ with a cost that is $m^2 d\log(1/\epsilon)$ where m is the dimension of the model, d the dimension of the space. The proposed approach consists of: first, computing the PSD model that satisfies the Fokker-Planck equation (or its fractional variant) associated with the SDE, up to error ϵ, and then sampling from the resulting PSD model. Assuming some regularity of the Fokker-Planck solution (i.e. β-times differentiability plus some geometric condition on its zeros) We obtain an algorithm that: (a) in the preparatory phase obtains a PSD model with L2 distance ϵ from the solution of the equation, with a model of dimension $m = \epsilon^{-(d+1)/(\beta-2s)}(\log(1/\epsilon))^{d+1}$ where $1/2 \leq s \leq 1$ is the fractional power to the Laplacian, and total computational complexity of $O(m^{3.5}\log(1/\epsilon))$ and then (b) for Fokker-Planck equation, it is able to produce i.i.d. samples with error ϵ in Wasserstein-1 distance, with a cost that is $O(d\epsilon^{-2(d+1)/\beta-2}\log(1/\epsilon)^{2d+3})$ per sample. This means that, if the probability associated with the SDE is somewhat regular, i.e. $\beta \geq 4d+2$, then the algorithm requires $O(\epsilon^{-0.88}\log(1/\epsilon)^{4.5d})$ in the preparatory phase, and $O(\epsilon^{-1/2}\log(1/\epsilon)^{2d+2})$ for each sample. Our results suggest that as the true solution gets smoother, we can circumvent the curse of dimensionality without requiring any sort of convexity. More information can be found in [14].

7.20 Generalization Guarantees via Algorithm-dependent Rademacher Complexity

Algorithm- and data-dependent generalization bounds are required to explain the generalization behavior of modern machine learning algorithms. In this context, there exists information theoretic generalization bounds that involve (various forms of) mutual information, as well as bounds based on hypothesis set stability. We propose a conceptually related, but technically distinct complexity measure to control generalization error, which is the empirical Rademacher complexity of an algorithm- and data-dependent hypothesis class. Combining standard properties of Rademacher complexity with the convenient structure of this class, we are able to (i) obtain novel bounds based on the finite fractal dimension, which (a) extend previous fractal dimension-type bounds from continuous to finite hypothesis classes, and (b) avoid a mutual information term that was required in prior work; (ii) we greatly simplify the proof of a recent dimension-independent generalization bound for stochastic gradient descent; and (iii) we easily recover results for VC classes and compression schemes, similar to approaches based on conditional mutual information. More information can be found in [17].

7.21 Generalization Bounds using Data-Dependent Fractal Dimensions

Providing generalization guarantees for modern neural networks has been a crucial task in statistical learning. Recently, several studies have attempted to analyze the generalization error in such settings by using tools from fractal geometry. While these works have successfully introduced new mathematical tools to apprehend generalization, they heavily rely on a Lipschitz continuity assumption, which in general does not hold for neural networks and might make the bounds vacuous. In this work, we address this issue and prove fractal geometry-based generalization bounds without requiring any Lipschitz assumption. To achieve this goal, we build up on a classical covering argument in learning theory and introduce a data-dependent fractal dimension. Despite introducing a significant amount of technical complications, this new notion lets us control the generalization error (over either fixed or random hypothesis spaces) along with certain mutual information (MI) terms. To provide a clearer interpretation to the newly introduced MI terms, as a next step, we introduce a notion of 'geometric stability' and link our bounds to the prior art. Finally, we make a rigorous connection between the proposed data-dependent dimension and topological data analysis tools, which then enables us to compute the dimension in a numerically efficient way. We support our theory with experiments conducted on various settings. More information can be found in [4].

7.22 Algorithmic Stability of Heavy-Tailed SGD with General Loss Functions

Heavy-tail phenomena in stochastic gradient descent (SGD) have been reported in several empirical studies. Experimental evidence in previous works suggests a strong interplay between the heaviness of the tails and generalization behavior of SGD. To address this empirical phenomena theoretically, several works have made strong topological and statistical assumptions to link the generalization error to heavy tails. Very recently, new generalization bounds have been proven, indicating a non-monotonic relationship between the generalization error and heavy tails, which is more pertinent to the reported empirical observations. While these bounds do not require additional topological assumptions given that SGD can be modeled using a heavy-tailed stochastic differential equation (SDE), they can only apply to simple quadratic problems. In this work, we build on this line of research and develop generalization bounds for a more general class of objective functions, which includes non-convex functions as well. Our approach is based on developing Wasserstein stability bounds for heavy-tailed SDEs and their discretizations, which we then convert to generalization bounds. Our results do not require any nontrivial assumptions; yet, they shed more light to the empirical observations, thanks to the generality of the loss functions. More information can be found in [15].

7.23 Algorithmic Stability of Heavy-Tailed Stochastic Gradient Descent on Least Squares

Heavy-tail phenomena in stochastic gradient descent (SGD) have been reported in several empirical studies. Experimental evidence in previous works suggests a strong interplay between the heaviness of the tails and generalization behavior of SGD. To address this empirical phenomena theoretically, several works have made strong topological and statistical assumptions to link the generalization error to heavy tails. Very recently, new generalization bounds have been proven, indicating a non-monotonic relationship between the generalization error and heavy tails, which is more pertinent to the reported empirical observations. While these bounds do not require additional topological assumptions given that SGD can be modeled using a heavy-tailed stochastic differential equation (SDE), they can only apply to simple quadratic problems. In this work, we build on this line of research and develop generalization bounds for a more general class of objective functions, which includes non-convex functions as well. Our approach is based on developing Wasserstein stability bounds for heavy-tailed SDEs and their discretizations, which we then convert to generalization bounds. Our results do not require any nontrivial assumptions; yet, they shed more light to the empirical observations, thanks to the generality of the loss functions. More information can be found in [13].

7.24 Cyclic and Randomized Stepsizes Invoke Heavier Tails in SGD than Constant Stepsize

Cyclic and randomized stepsizes are widely used in the deep learning practice and can often outperform standard stepsize choices such as constant stepsize in SGD. Despite their empirical success, not much is currently known about when and why they can theoretically improve the generalization performance. We consider a general class of Markovian stepsizes for learning, which contain i.i.d. random stepsize, cyclic stepsize as well as the constant stepsize as special cases, and motivated by the literature which shows that heaviness of the tails (measured by the so-called "tail-index") in the SGD iterates is correlated with generalization, we study tail-index and provide a number of theoretical results that demonstrate how the tail-index varies on the stepsize scheduling. Our results bring a new understanding of the benefits of cyclic and randomized stepsizes compared to constant stepsize in terms of the tail behavior. We illustrate our theory on linear regression experiments and show through deep learning experiments that Markovian stepsizes can achieve even a heavier tail and be a viable alternative to cyclic and i.i.d. randomized stepsize rules. More information can be found in [5].

7.25 An Oblivious Stochastic Composite Optimization Algorithm for Eigenvalue Optimization Problems

In this work, we revisit the problem of solving large-scale semidefinite programs using randomized first-order methods and stochastic smoothing. We introduce two oblivious stochastic mirror descent algorithms based on a complementary composite setting. One algorithm is designed for non-smooth objectives, while an accelerated version is tailored for smooth objectives. Remarkably, both algorithms work without prior knowledge of the Lipschitz constant or smoothness of the objective function. For the non-smooth case with $\mathscr{M}-$bounded oracles, we prove a convergence rate of $O(\mathscr{M}/\sqrt{T})$. For the L-smooth case with a feasible set bounded by D, we derive a convergence rate of $O(L^2D^2/(T^2\sqrt{T})+(D_0^2+\sigma^2)/\sqrt{T})$, where D_0 is the starting distance to an optimal solution, and σ^2 is the stochastic oracle variance. These rates had only been obtained so far by either assuming prior knowledge of the Lipschitz constant or the starting distance to an optimal solution. We further show how to extend our framework to relative scale and demonstrate the efficiency and robustness of our methods on large scale semidefinite programs.

7.26 Vision Transformers, a new approach for high-resolution and large-scale mapping of canopy heights

Accurate and timely monitoring of forest canopy heights is critical for assessing forest dynamics, biodiversity, carbon sequestration as well as forest degradation and deforestation. Recent advances in deep learning techniques, coupled with the vast amount of spaceborne remote sensing data offer an unprecedented opportunity to map canopy height at high spatial and temporal resolutions. Current techniques for wall-to-wall canopy height mapping correlate remotely sensed 2D information from optical and radar sensors to the vertical structure of trees using LiDAR measurements. While studies using deep learning algorithms have shown promising performances for the accurate mapping of canopy heights, they have limitations due to the type of architectures and loss functions employed. Moreover, mapping canopy heights over tropical forests remains poorly studied, and the accurate height estimation of tall canopies is a challenge due to signal saturation from optical and radar sensors, persistent cloud covers and sometimes the limited penetration capabilities of LiDARs. Here, we map heights at 10 m resolution across the diverse landscape of Ghana with a new vision transformer (ViT) model optimized concurrently with a classification (discrete) and a regression (continuous) loss function. This model achieves better accuracy than previously used convolutional based approaches (ConvNets) optimized with only a continuous loss function. The ViT model results show that our proposed discrete/continuous loss significantly increases the sensitivity for very tall trees (i.e., > 35m), for which other approaches show saturation effects. The height maps generated by the ViT also have better ground sampling distance and better sensitivity to sparse vegetation in comparison to a convolutional model. Our ViT model has a RMSE of 3.12m in comparison to a reference dataset while the ConvNet model has a RMSE of 4.3m.

8 Bilateral contracts and grants with industry

8.1 Bilateral grants with industry

- Alexandre d'Aspremont, Francis Bach, Martin Jaggi (EPFL): Google Focused award.

- Francis Bach: Gift from Facebook AI Research.

9 Partnerships and cooperations

9.1 International initiatives

9.1.1 Associate Teams in the framework of an Inria International Lab or in the framework of an Inria International Program

FOAM

Title: First-Order Accelerated Methods for Machine Learning.

Duration: 2020 -> present

Coordinator: Cristobal Guzman (crguzmanp@mat.uc.cl)

Partners: Pontificia Universidad Católica de Chile Santiago (Chili)

Inria contact: Alexandre d'Aspremont

Summary: Our main interest is to investigate novel and improved convergence results for first-order iterative methods for saddle-points, variational inequalities and fixed points, under the lens of PEP. Our interest in improving first-order methods is also deeply related with applications in machine learning. Particularly in sparsity-oriented inverse problems, optimization methods are the workhorse for state of the art results. On some of these problems, a set of new hypothesis and theoretical results shows improved complexity bounds for problems with good recovery guarantees and we plan to extend these new performance bounds to the variational framework.

4TUNE

Title: Adaptive, Efficient, Provable and Flexible Tuning for Machine Learning

Duration: 020 ->

Coordinator: Peter Grünwald (pdg@cwi.nl)

Partners: CWI

Inria contact: Adrien Taylor

Summary: The long-term goal of 4TUNE is to push adaptive machine learning to the next level. We aim to develop refined methods, going beyond traditional worst-case analysis, for exploiting structure in the learning problem at hand. We will develop new theory and design sophisticated algorithms for the core tasks of statistical learning and individual sequence prediction. We are especially interested in understanding the connections between these tasks and developing unified methods for both. We will also investigate adaptivity to non-standard patterns encountered in embedded learning tasks, in particular in iterative equilibrium computations.

9.1.2 Visits of international scientists

Inria International Chair Laurent El Ghaoui (U.C. Berkeley)

Other international visits to the team Antônio Horta Ribeiro (University of Upsalla)
Steffen Grunewalder (University of Newcastle)

9.2 European initiatives

9.2.1 Horizon Europe

DYNASTY DYNASTY project on cordis.europa.eu

Title: Dynamics-Aware Theory of Deep Learning

Duration: From October 1, 2022 to September 30, 2027

Partners:

- INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET AUTOMATIQUE (INRIA), France

Inria contact: Umut SIMSEKLI

Coordinator:

Summary: The recent advances in deep learning (DL) have transformed many scientific domains and have had major impacts on industry and society. Despite their success, DL methods do not obey most of the wisdoms of statistical learning theory, and the vast majority of the current DL techniques mainly stand as poorly understood black-box algorithms.

Even though DL theory has been a very active research field in the past few years, there is a significant gap between the current theory and practice: (i) the current theory often becomes vacuous for models with large number of parameters (which is typical in DL), and (ii) it cannot capture the interaction between data, architecture, training algorithm and its hyper-parameters, which can have drastic effects on the overall performance. Due to this lack of theoretical understanding, designing new DL systems has been dominantly performed by ad-hoc, 'trial-and-error' approaches.

The main objective of this proposal is to develop a mathematically sound and practically relevant theory for DL, which will ultimately serve as the basis of a software library that provides practical tools for DL practitioners. In particular, (i) we will develop error bounds that closely reflect the true empirical performance, by explicitly incorporating the dynamics aspect of training, (ii) we will develop new model selection, training, and compression algorithms with reduced time/memory/storage complexity, by exploiting the developed theory.

To achieve the expected breakthroughs, we will develop a novel theoretical framework, which will enable tight analysis of learning algorithms in the lens of dynamical systems theory. The outcomes will help relieve DL from being a black-box system and avoid the heuristic design process. We will produce comprehensive open-source software tools adapted to all popular DL libraries, and test the developed algorithms on a wide range of real applications arising in computer vision, audio/music/natural language processing.

9.2.2 H2020 projects

SEQUOIA SEQUOIA project on cordis.europa.eu

Title: Robust algorithms for learning from modern data

Duration: From September 1, 2017 to August 31, 2023

Partners:

- INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET AUTOMATIQUE (INRIA), France

Inria contact: Francis BACH

Coordinator:

Summary: Machine learning is needed and used everywhere, from science to industry, with a growing impact on many disciplines. While first successes were due at least in part to simple supervised learning algorithms used primarily as black boxes on medium-scale problems, modern data pose new challenges. Scalability is an important issue of course: with large amounts of data, many current problems far exceed the capabilities of existing algorithms despite sophisticated computing architectures. But beyond this, the core classical model of supervised machine learning, with the usual assumptions of independent and identically distributed data, or well-defined features, outputs and loss functions, has reached its theoretical and practical limits.

Given this new setting, existing optimization-based algorithms are not adapted. The main objective of this proposal is to push the frontiers of supervised machine learning, in terms of (a) scalability to data with massive numbers of observations, features, and tasks, (b) adaptability to modern computing environments, in particular for parallel and distributed processing, (c) provable adaptivity and robustness to problem and hardware specifications, and (d) robustness to non-convexities inherent in machine learning problems.

To achieve the expected breakthroughs, we will design a novel generation of learning algorithms amenable to a tight convergence analysis with realistic assumptions and efficient implementations. They will help transition machine learning algorithms towards the same wide-spread robust use as numerical linear algebra libraries. Outcomes of the research described in this proposal will include algorithms that come with strong convergence guarantees and are well-tested on real-life benchmarks coming from computer vision, bioinformatics, audio processing and natural language processing. For both distributed and non-distributed settings, we will release open-source software, adapted to widely available computing platforms.

REAL REAL project on cordis.europa.eu

Title: Reliable and cost-effective large scale machine learning

Duration: From April 1, 2021 to March 31, 2026

Partners:

- INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET AUTOMATIQUE (INRIA), France

Inria contact: Alessandro Rudi

Coordinator:

Summary: In the last decade, machine learning (ML) has become a fundamental tool with a growing impact in many disciplines, from science to industry. However, nowadays, the scenario is changing: data are exponentially growing compared to the computational resources (post Moore's law era), and ML algorithms are becoming crucial building blocks in complex systems for decision making, engineering, science. Current machine learning is not suitable for the new scenario, both from a theoretical and a practical viewpoint: (a) the lack of cost-effectiveness of the algorithms impacts directly the economic/energetic costs of large scale ML, making it barely affordable by universities or research institutes; (b) the lack of reliability of the predictions affects critically the safety of the systems where ML is employed. To deal with the challenges posed by the new scenario, REAL will lay the foundations of a solid theoretical and algorithmic framework for reliable and cost-effective large scale machine learning on modern computational architectures. In particular, REAL will extend the classical ML framework to provide algorithms with two additional guarantees: (a) the predictions will be reliable, i.e., endowed with explicit bounds on their uncertainty guaranteed by the theory; (b) the algorithms will be cost-effective, i.e., they will be naturally adaptive to the new architectures and will provably achieve the desired reliability and accuracy level, by using minimum possible computational resources. The algorithms resulting from REAL will be released as open-source libraries for distributed and multi-GPU settings, and their effectiveness will be extensively tested on key benchmarks from computer vision, natural language processing, audio processing, and bioinformatics. The methods and the techniques developed in this project will help machine learning to take the next step and become a safe, effective, and fundamental tool in science and engineering for large scale data problems.

NN-OVEROPT NN-OVEROPT project on cordis.europa.eu

Title: Neural Network : An Overparametrization Perspective

Duration: From November 1, 2021 to October 31, 2024

Partners:

- INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET AUTOMATIQUE (INRIA), France
- THE BOARD OF TRUSTEES OF THE UNIVERSITY OF ILLINOIS (UNIVERSITY OF ILLINOIS), United States

Inria contact: Francis Bach

Coordinator:

Summary: In recent times, overparametrized models where the number of model parameters far exceeds the number of training samples available are the methods of choice for learning problems and neural networks are amongst the most popular overparametrized methods used heavily in practice. It has been discovered recently that overparametrization surprisingly improves the optimization landscape of a complex non-convex problem, i.e., the training of neural networks, and also has positive effects on the generalization performance. Despite improved empirical performance of overparametrized models like neural networks, the theoretical understanding of these models is quite limited which hinders the progress of the field in the right direction. Any progress in the understanding of the optimization as well as generalization aspects for theses complex models especially neural networks will lead to big technical advancement in the field of machine learning and artificial intelligence. During the Marie Sklodowska-Curie Actions Individual Fellowship-Global Fellowship (MSCA-IF-GF), I plan to study the optimization problem arising while training overparametrized neural networks and generalization in overparametrized neural networks. The end goal for this project is to provide better theoretical understanding of the optimization landscape while training overparametrized models as a result of which to provide better optimization algorithms for training as well as to study the universal approximation guarantees of overparametrized models. We also aim to study the implicit bias induced by optimization algorithms while training overparametrized complex models. To achieve the objective discussed above, I will be using tools from traditional optimization theory, statistical learning theory, gradient flows, as well as from statistical physics.

10 Dissemination

10.1 Promoting scientific activities

10.1.1 Journal

Member of the editorial boards

- A. d'Aspremont, Section Editor, SIAM Journal on the Mathematics of Data Science.
- F. Bach, co-editor-in-chief, Journal of Machine Learning Research
- F. Bach: Series Editor, Adaptive Computation and Machine Learning, MIT Press, since 2016.

Reviewer - reviewing activities

- Adrien Taylor: Mathematical Programming.
- Adrien Taylor: SIAM Journal on Optimization.
- Adrien Taylor: Transactions on Automatic Control.
- Adrien Taylor: Journal of Optimization Theory and Applications.

10.1.2 Invited talks

- Alessandro Rudi: StatML Summer School in Causality and Statistical Learning, Oxford, July 2023
- Alessandro Rudi: 7th London Symposium on Information Theory, University College of London, May 2023
- Alessandro Rudi: PhD school on PSD Models, Scuola Normale Superiore, Italy, Jan 2023
- Alessandro Rudi: Optimization and Statistical Learning Workshop, Les Houches, France, Jan 2023
- Francis Bach: Optimization and Statistical Learning workshop, les Houches. January 2023
- F. Bach: FOCM conference, June 2023
- F. Bach: ICIAM conference, plenary talk, Tokyo, June 2023
- F. Bach: GSI conference, keynote speaker, Saint Malo, September 2023
- F. Bach: NCCR Symposium, Zurich, September 2023
- F. Bach: GDR RSD Summer School on Distributed Learning, Lyon, September 2023
- F. Bach: "IA et commandement militaire" Day, September 2023
- F. Bach: POP23 - Future Trends in Polynomial OPtimization, Toulouse, October 2023
- U. Simsekli: MIT, Boston, July 2023
- U. Simsekli: Flatiron Institute, New York, July 2023
- U. Simsekli: Berkeley University, Berkeley, July 2023
- M. Jordan: Distinguished Lecture, Indian Institute of Science, Bangalore, India, July 2023
- M. Jordan: Turing Lectures, International Centre for Theoretical Science, Bangalore, India, July 2023
- M. Jordan: Leonardo da Vinci Lectures (Lezioni Leonardesche), Milan, October 2023

10.1.3 Scientific expertise

- F. Bach: President of the board of ICML (until July 2023)
- F. Bach: Member of the Scientific Council of the Société Informatique de France, since 2022.

10.2 Teaching - Supervision - Juries

10.2.1 Teaching

- Master: Alexandre d'Aspremont, Optimisation convexe: modélisation, algorithmes et applications cours magistraux 21h (2011-Present), Master M2 MVA, ENS PS.
- Master : Francis Bach, Learning theory from first principles, 27h, Master M2 MASH, Université Paris Dauphine PSL, France.
- Master: Alessandro Rudi, Umut Simsekli. Introduction to Machine Learning, 52h, L3, ENS, Paris.
- Master: Alessandro Rudi, Kernel Methods 10h, Master M2 MVA, ENS PS.
- Master: Adrien Taylor. Convex Optimization, 21h, M1, ENS, Paris.
- Master : Umut Simsekli. Deep Learning, 21h, M2, Ecole Polytechnique, Palaiseau, France

10.2.2 Supervision

- PhD in progress: Benjamin Dupuis, current PhD Student, supervised by Umut Simsekli
- PhD in progress: Dario Shariatian, current PhD Student, co-supervised by Umut Simsekli and Alain Durmus
- PhD in progress: Andrea Basteri, current PhD Student, co-supervised by Alessandro Rudi and Fancis Bach
- PhD in progress: Pierre Boudart, current PhD Student, co-supervised by Alessandro Rudi, Pierre Gaillard and Alexandre d'Aspremont
- PhD in progress: Theophile Cantelobre, current PhD student, co-supervised by Alessandro Rudi, Benjamin Guedj
- PhD in progress: Gaspard Beugnot, current PhD student, co-supervised by Alessandro Rudi, Julien Mairal
- PhD in progress: Antoine Bambade supervised by Jean Ponce (WILLOW), Justin Carpentier (WILLOW), and Adrien Taylor.
- PhD in progress: Baptiste Goujaud supervised by Eric Moulines (École Polytechnique), Aymeric Dieuleveut (École Polytechnique), and Adrien Taylor.
- PhD in progress: Céline Moucer supervised by Francis Bach and Adrien Taylor.
- PhD in progress: Bertille Follain supervised by F. Bach and U. Simsekli.
- PhD in progress: Marc Lambert supervised by F. Bach and S. Bonnabel.
- PhD in progress: Ivan Lerner, co-advised with Anita Burgun et Antoine Neuraz.
- PhD in progress: Lawrence Stewart, co-advised by Francis Bach and Jean-Philippe Vert.
- PhD in progress: Gautier Izacard, co-advised by Alexandre d'Aspremont and Edouard Grave (Meta).
- PhD in progress: Cle'ment Lezane, co-advised by Alexandre d'Aspremont and Cristobal Guzman.
- PhD in progress: Sarah Brood, co-advised by Alexandre d'Aspremont and Philippe Ciais.
- PhD in progress: Arthur Calvi, co-advised by Alexandre d'Aspremont and Philippe Ciais.
- PhD in progress: Benjamin Dubois-Taine, co-advised by Alexandre d'Aspremont and Alessandro Rudi.
- PhD defended: Rémi Jezequel, April 18, 2023

10.2.3 Juries

- F. Bach: PHD examiner for Paul Youssef (Laboratoire d'Informatique de Grenoble)
- A. d'Aspremont: PHD examiner for Hippolyte Labarrière (U. de Toulouse)
- A. d'Aspremont: PHD examiner for Lucie Neirac (IP Paris)

10.3 Popularization

10.3.1 Education

A. d'Aspremont: Geospatial data and AI, Step in Stem, EJM Paris.

10.3.2 Interventions

A. d'Aspremont: rencontres de l'avenir, St Raphaël.

11 Scientific production

11.1 Major publications

[1] A. Askari, A. d'Aspremont and L. E. Ghaoui. 'Approximation Bounds for Sparse Programs'. In: *SIAM Journal on Mathematics of Data Science* 4.2 (1st June 2022), pp. 514–530. DOI: 10.1137/21M1398677. URL: https://hal.archives-ouvertes.fr/hal-03165622.

[2] T. Cantelobre, C. Ciliberto, B. Guedj and A. Rudi. *Measuring dissimilarity with diffeomorphism invariance*. 24th Feb. 2022. DOI: 10.48550/arXiv.2202.05614. URL: https://hal.inria.fr/hal-03573479.

[3] R.-A. Dragomir, A. Taylor, A. d'Aspremont and J. Bolte. 'Optimal Complexity and Certification of Bregman First-Order Methods'. In: *Mathematical Programming* 194.1 (1st July 2022), pp. 41–83. DOI: 10.1007/s10107-021-01618-1. URL: https://hal.inria.fr/hal-02384167.

[4] B. Dupuis, G. Deligiannidis and U. Şimşekli. 'Generalization Bounds using Data-Dependent Fractal Dimensions'. In: *Proceedings of Machine Learning Research*. International Conference on Machine Learning (ICML 2023). Honolulu, United States, 10th July 2023. URL: https://inria.hal.science/hal-04438550.

[5] M. Gürbüzbalaban, Y. Hu, U. Şimşekli and L. Zhu. 'Cyclic and Randomized Stepsizes Invoke Heavier Tails in SGD than Constant Stepsize'. In: *Transactions on Machine Learning Research Journal* (2023). URL: https://inria.hal.science/hal-04478948.

[6] L. Hodgkinson, U. Şimşekli, R. Khanna and M. W. Mahoney. 'Generalization Bounds using Lower Tail Exponents in Stochastic Optimizers'. In: International Conference on Machine Learning. Baltimore, United States, 2022. URL: https://hal.inria.fr/hal-03935798.

[7] S. Kolouri, K. Nadjahi, S. Shahrampour and U. Simsekli. 'Generalized Sliced Probability Metrics'. In: ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Singapore, Singapore: IEEE, 23rd May 2022, pp. 4513–4517. DOI: 10.1109/ICASSP43922.2022.9746016. URL: https://hal.inria.fr/hal-03935833.

[8] T. Lauvaux, C. Giron, M. Mazzolini, A. d'Aspremont, R. Duren, D. Cusworth, D. Shindell and P. Ciais. 'Global assessment of oil and gas methane ultra-emitters'. In: *Science* 375.6580 (4th Feb. 2022), pp. 557–561. DOI: 10.1126/science.abj4351. URL: https://hal.archives-ouvertes.fr/hal-03565371.

[9] S. H. Lim, Y. Wan and U. Şimşekli. 'Chaotic Regularization and Heavy-Tailed Limits for Deterministic Gradient Descent'. In: Advances in Neural Processing Systems. New Orleans, United States, 2022. URL: https://hal.inria.fr/hal-03935819.

[10] U. Marteau-Ferey, F. Bach and A. Rudi. 'Non-parametric Models for Non-negative Functions'. working paper or preprint. July 2020. URL: https://hal.inria.fr/hal-02891640.

[11] S. Park, U. Şimşekli and M. A. Erdogdu. 'Generalization Bounds for Stochastic Gradient Descent via Localized ε-Covers'. In: Advances in Neural Processing Systems. Baltimore, United States, 19th Sept. 2022. URL: https://hal.inria.fr/hal-03935856.

[12] K. L. Pavasovic, A. Durmus and U. Simsekli, eds. *Approximate Heavy Tails in Offline (Multi-Pass) Stochastic Gradient Descent.* Advances in Neural Information Processing Systems. 27th Oct. 2023. URL: https://inria.hal.science/hal-04478942.

[13] A. Raj, M. Barsbey, M. Gürbüzbalaban, L. Zhu and U. Şimşekli. 'Algorithmic Stability of Heavy-Tailed Stochastic Gradient Descent on Least Squares'. In: Algorithmic Learning Theory. Singapore, Singapore, 2023. URL: https://inria.hal.science/hal-04478947.

[14] A. Raj, U. Şimşekli and A. Rudi, eds. *Efficient Sampling of Stochastic Differential Equations with Positive Semi-Definite Models.* Advances in Neural Information Processing Systems. 2023. URL: https://inria.hal.science/hal-04478943.

[15] A. Raj, L. Zhu, M. Gürbüzbalaban and U. Şimşekli. 'Algorithmic Stability of Heavy-Tailed SGD with General Loss Functions'. In: International Conference on Machine Learning. Honolulu, United States, 2023. URL: https://inria.hal.science/hal-04478946.

[16] V. Roulet and A. D'Aspremont. 'Sharpness, Restart and Acceleration'. In: *SIAM Journal on Optimization* 30.1 (Oct. 2020), pp. 262–289. DOI: 10.1137/18M1224568. URL: https://hal.archives-ouvertes.fr/hal-02983236.

[17] S. Sachs, T. van Erven, L. Hodgkinson, R. Khanna and U. Simsekli. 'Generalization Guarantees via Algorithm-dependent Rademacher Complexity'. In: Conference on Learning Theory. Bangalore (Virtual event), India, 4th July 2023. URL: https://inria.hal.science/hal-04478945.

[18] M. Sefidgaran, A. Gohari, G. Richard and U. Şimşekli. 'Rate-Distortion Theoretic Generalization Bounds for Stochastic Learning Algorithms'. In: COLT 2022 - 35th Annual Conference on Learning Theory. Vol. 178. Proceedings of Machine Learning Research. London, United Kingdom, 2nd July 2022. URL: https://hal.telecom-paris.fr/hal-03759597.

[19] P. Viallard, M. Haddouche, U. Şimşekli and B. Guedj. 'Learning via Wasserstein-Based High Probability Generalisation Bounds'. In: NeurIPS 2023 - Thirty-seventh Conference on Neural Information Processing Systems. New Orleans, United States, 7th June 2023. DOI: 10.48550/arXiv.2306.04375. URL: https://hal.science/hal-04121624.

[20] B. Woodworth, F. Bach and A. Rudi. *Non-Convex Optimization with Certificates and Fast Rates Through Kernel Sums of Squares.* 8th Apr. 2022. DOI: 10.48550/arXiv.2204.04970. URL: https://hal.inria.fr/hal-03635236.

[21] L. Zhu, M. Gurbuzbalaban, A. Raj and U. Simsekli, eds. *Uniform-in-Time Wasserstein Stability Bounds for (Noisy) Stochastic Gradient Descent.* Advances in Neural Information Processing Systems. 2023. URL: https://inria.hal.science/hal-04478941.

11.2 Publications of the year

International journals

[22] F. Bach and A. Rudi. 'Exponential convergence of sum-of-squares hierarchies for trigonometric polynomials'. In: *SIAM Journal on Optimization* (2023). DOI: 10.1137/22m1540818. URL: https://hal.science/hal-03843458.

[23] B. Goujaud, A. Dieuleveut and A. Taylor. 'Counter-Examples in First-Order Optimization: A Constructive Approach'. In: *IEEE Control Systems Letters* 7 (2023), pp. 2485–2490. DOI: 10.1109/LCSYS.2023.3286277. URL: https://hal.science/hal-04384238.

[24] M. Lambert, S. Bonnabel and F. Bach. 'The limited-memory recursive variational Gaussian approximation (L-RVGA)'. In: *Statistics and Computing* 33.70 (2023). DOI: 10.1007/s11222-023-10239-x. URL: https://inria.hal.science/hal-03501920.

[25] M. Lambert, S. Bonnabel and F. Bach. 'Variational Gaussian approximation of the Kushner optimal filter'. In: *Lecture Notes in Computer Science* (1st Aug. 2023). DOI: 10.1007/978-3-031-38271-0_39. URL: https://hal.science/hal-04218385.

[26] A. Taylor and Y. Drori. 'An optimal gradient method for smooth convex minimization'. In: *Mathematical Programming, Series A* 199.1-2 (May 2023), pp. 557–594. DOI: 10.1007/s10107-022-01839-y. URL: https://inria.hal.science/hal-03154583.

[27] P. Viallard, P. Germain, A. Habrard and E. Morvant. 'A General Framework for the Practical Disintegration of PAC-Bayesian Bounds'. In: *Machine Learning* (2023). URL: https://hal.science/hal-03143025.

International peer-reviewed conferences

[28] G. Beugnot, J. Mairal and A. Rudi. 'GloptiNets: Scalable Non-Convex Optimization with Certificates'. In: NeurIPS 2023 - 37th Conference on Neural Information Processing Systems. New Orleans, United States, Dec. 2023, pp. 1–21. URL: https://inria.hal.science/hal-04138843.

[29] E. Gorbunov, A. Taylor, S. Horváth and G. Gidel. 'Convergence of Proximal Point and Extragradient-Based Methods Beyond Monotonicity: the Case of Negative Comonotonicity'. In: *Proceedings of the 40th International Conference on Machine Learning, PMLR 202:11614-11641, 2023.* ICML 2023 - 40th International Conference on Machine Learning. Honolulu, Hawai, United States, 23rd July 2023. DOI: 10.48550/arXiv.2210.13831. URL: https://hal.science/hal-04384208.

[30] B. Goujaud, A. Dieuleveut and A. Taylor. 'On Fundamental Proof Structures in First-Order Optimization'. In: Conference on Decision and Control, Tutorial sessions. Marina Bay Sands, Singapore: arXiv, 2023. DOI: 10.48550/arXiv.2310.02015. URL: https://hal.science/hal-04384178.

[31] K. L. Pavasovic, A. Durmus and U. Simsekli. 'Approximate Heavy Tails in Offline (Multi-Pass) Stochastic Gradient Descent'. In: Neural Information Processing Systems (NeurIPS), Spotlight Presentation, 2023. New Orleans (LA), United States, 27th Oct. 2023. URL: https://hal.science/hal-04271020.

[32] L. Stewart, F. Bach, Q. Berthet and J.-P. Vert. 'Regression as Classification: Influence of Task Formulation on Neural Network Features'. In: AISTATS 2023 - 26th International Conference on Artificial Intelligence and Statistics. Vol. 206. Valence, Spain, 2023. URL: https://hal.science/hal-03846706.

[33] P. Viallard, M. Haddouche, U. Şimşekli and B. Guedj. 'Learning via Wasserstein-Based High Probability Generalisation Bounds'. In: NeurIPS 2023 Workshop on Optimal Transport and Machine Learning (OTML'23). New Orleans, United States, 16th Dec. 2023. URL: https://hal.science/hal-04273718.

[34] P. Viallard, M. Haddouche, U. Şimşekli and B. Guedj. 'Learning via Wasserstein-Based High Probability Generalisation Bounds'. In: NeurIPS 2023 - Thirty-seventh Conference on Neural Information Processing Systems. New Orleans, United States, 7th June 2023. DOI: 10.48550/arXiv.2306.04375. URL: https://hal.science/hal-04121624.

Conferences without proceedings

[35] P. Viallard. 'Bornes de généralisation : quand l'information mutuelle rencontre les bornes PAC-Bayésiennes et désintégrées'. In: CAp 2023 - Conférence sur l'Apprentissage Automatique. Strasbourg, France, 3rd July 2023. URL: https://hal.science/hal-04093184.

Reports & preprints

[36] F. Bach. *High-dimensional analysis of double descent for linear regression with random projections.* 28th Feb. 2023. URL: https://hal.science/hal-04008311.

[37] F. Bach. *On the relationship between multivariate splines and infinitely-wide neural networks.* 6th Feb. 2023. URL: https://hal.science/hal-03974669.

[38] F. Bach. *Sum-of-Squares Relaxations for Information Theory and Variational Inference.* 15th Sept. 2023. URL: https://hal.science/hal-03703475.

[39] F. Bach. *Sum-of-squares relaxations for polynomial min-max problems over simple sets.* 25th June 2023. URL: https://hal.science/hal-04140288.

[40] A. Bambade, F. Schramm, S. E. Kazdadi, S. Caron, A. Taylor and J. Carpentier. *Companion Report of PROXQP: an Efficient and Versatile Quadratic Programming Solver for Real-Time Robotics Applications and Beyond*. INRIA, Sept. 2023. URL: https://inria.hal.science/hal-04196897.

[41] A. Bambade, F. Schramm, S. E. Kazdadi, S. Caron, A. Taylor and J. Carpentier. *PROXQP: an Efficient and Versatile Quadratic Programming Solver for Real-Time Robotics Applications and Beyond*. 1st Sept. 2023. URL: https://inria.hal.science/hal-04198663.

[42] A. Bambade, F. Schramm, A. Taylor and J. Carpentier. *QPLayer: efficient differentiation of convex quadratic optimization*. 19th June 2023. URL: https://inria.hal.science/hal-04133055.

[43] E. Berta, F. Bach and M. Jordan. *Classifier Calibration with ROC-Regularized Isotonic Regression*. 20th Nov. 2023. URL: https://hal.science/hal-04295601.

[44] R. Bonalli and A. Rudi. *Non-Parametric Learning of Stochastic Differential Equations with Fast Rates of Convergence*. 24th May 2023. URL: https://hal.science/hal-04381810.

[45] I. Fayad, P. Ciais, M. Schwartz, J.-P. Wigneron, N. Baghdadi, A. de Truchis, A. d'Aspremont, F. Frappart, S. Saatchi, A. Pellissier-Tanon and H. Bazzi. *Vision Transformers, a new approach for high-resolution and large-scale mapping of canopy heights*. 22nd Apr. 2023. URL: https://hal.science/hal-04230906.

[46] B. Follain, U. Şimşekli and F. Bach. *Nonparametric Linear Feature Learning in Regression Through Regularisation*. 24th July 2023. URL: https://hal.science/hal-04170331.

[47] B. Goujaud, A. Taylor and A. Dieuleveut. *Provable non-accelerations of the heavy-ball method*. 2023. DOI: 10.48550/arXiv.2307.11291. URL: https://hal.science/hal-04384188.

[48] S. D. Gupta, R. Freund, X. A. Sun and A. Taylor. *Nonlinear conjugate gradient methods: worst-case convergence rates via computer-assisted analyses*. 2023. DOI: 10.48550/arXiv.2301.01530. URL: https://hal.science/hal-04384219.

[49] D. Holzmüller and F. Bach. *Convergence Rates for Non-Log-Concave Sampling and Log-Partition Estimation*. 6th Mar. 2023. URL: https://hal.science/hal-04018103.

[50] Z. Kobeissi and F. Bach. *Temporal Difference Learning with Continuous Time and State in the Stochastic Setting*. 2nd June 2023. URL: https://inria.hal.science/hal-03574645.

[51] F. Léger and P.-C. Aubin-Frankowski. *Gradient descent with a general cost*. 14th Dec. 2023. URL: https://hal.science/hal-04344054.

[52] C. Lezane, C. Guzmán and A. d'Aspremont. *An Oblivious Stochastic Composite Optimization Algorithm for Eigenvalue Optimization Problems*. 30th June 2023. URL: https://hal.science/hal-04230909.

[53] S. Martin, F. Bach and G. Biroli. *On the Impact of Overparameterization on the Training of a Shallow Neural Network in High Dimensions*. 4th Nov. 2023. URL: https://hal.science/hal-04270390.

[54] L. Montaut, Q. Le Lidec, A. Bambade, V. Petrík, J. Sivic and J. Carpentier. *Differentiable Collision Detection: a Randomized Smoothing Approach*. 14th Apr. 2023. URL: https://hal.science/hal-03780482.

[55] K. Nadjahi, V. de Bortoli, A. Durmus, R. Badeau and U. Şimşekli. *Approximate Bayesian computation with the sliced-Wasserstein distance*. 16th Jan. 2024. DOI: 10.1109/icassp40776.2020.9054735. URL: https://hal.science/hal-03945515.

[56] M. Upadhyaya, S. Banert, A. Taylor and P. Giselsson. *Automated tight Lyapunov analysis for first-order methods*. 2023. DOI: 10.48550/arXiv.2302.06713. URL: https://hal.science/hal-04384212.

[57] B. Woodworth, K. Mishchenko and F. Bach. *Two Losses Are Better Than One: Faster Optimization Using a Cheaper Proxy*. Feb. 2023. URL: https://inria.hal.science/hal-03977083.