IN PARTNERSHIP WITH:
CNRS

Ecole normale supérieure de Paris

Activity Report 2018

Project-Team SIERRA

Statistical Machine Learning and Parsimony

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

RESEARCH CENTER
Paris

THEME
Optimization, machine learning and statistical methods

Table of contents

Project-Team SIERRA

Creation of the Team: 2011 January 01, updated into Project-Team: 2012 January 01

Keywords:

Computer Science and Digital Science:

A1.2.8. - Network security
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, Visitors, External Collaborators

Research Scientists

Francis Bach [Team leader, Inria, Senior Researcher, HDR]
Alexandre d'Aspremont [CNRS, Senior Researcher, HDR]
Pierre Gaillard [Inria, Researcher]
Alessandro Rudi [Inria, Starting Research Position]

Post-Doctoral Fellows

Lenaic Chizat [Inria, until Nov 2018]
Pierre Yves Massé [Université Technique de Prague, from Apr 2018]
Dmitrii Ostrovskii [Inria, from Feb 2018]
Adrien Taylor [Inria]

PhD Students

Remi Leblond [Inria, Researcher, until Aug 2018]
Dmitry Babichev [Inria]
Mathieu Barré [Ecole Normale Supérieure Paris, from Sep 2018]
Raphaël Berthier [Inria, from Oct 2018]
Anaël Bonneton [Ecole Normale Supérieure Paris]
Margaux Brégère [EDF]
Alexandre Défossez [Facebook]
Radu Alexandru Dragomir [Ecole polytechnique, from Sep 2018]
Thomas Kerdreux [Ecole polytechnique]
Gregoire Mialon [Inria, from Oct 2018]
Loucas Pillaud Vivien [Ministère de l'Ecologie, de l'Energie, du Développement durable et de la Mer]
Antoine Recanati [CNRS, until Sep 2018]
Damien Scieur [Inria, until Aug 2018]
Tatiana Shpakova [Inria]
Alex Nowak Vila [Inria, from Oct 2018]

Technical staff

Loïc Estève [Inria, from Apr 2018]
Hadrien Hendrikx [Inria, from Apr 2018 until Sep 2018]

Interns

Mathieu Barre [Inria, from Apr 2018 until Sep 2018]
Raphaël Berthier [Ecole Normale Supérieure Paris, until Sep 2018]
Vivien Cabannes [Univ Vincennes-Saint Denis, from Sep 2018]
Florentin Guth [Ecole Normale Supérieure Paris, from Feb 2018 until Mar 2018]
Remi Jézequel [Ecole Normale Supérieure Paris, from Oct 2018]
Ulysse Marteau Ferey [Ecole Normale Supérieure Paris, from Apr 2018]
Alex Nowak Vila [Inria, from Apr 2018 until Sep 2018]

Administrative Assistants

Helene Bessin Rousseau [Inria, from Mar 2018]
Sabrine Boumizy [Inria, until Feb 2018]
Sandrine Verges [Inria, until Jan 2018]

Visiting Scientists

Vijaya Bollapragada [Northwestern University, from Apr 2018 until Jul 2018]
Aaron Defazio [Facebook Research, until Feb 2018]
Gauthier Gidel [University of Montreal, Jan 2018]
Achintya Kundu [Ecole d'ingénieurs, from Jun 2018 until Aug 2018]
Gregoire Mialon [Inria, Sep 2018]
Sharan Vaswani [University of British Columbia, from Apr 2018 until Jul 2018]
Simon Lacoste-Julien [University of Montreal, Aug 2018]

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, text processing and neuro-imaging.

Machine learning is now a vast field of research and the team focuses on the following aspects: supervised learning (kernel methods, calibration), unsupervised learning (matrix factorization, statistical tests), parsimony (structured sparsity, theory and algorithms), and optimization (convex optimization, bandit learning). These four research axes are strongly interdependent, and the interplay between them is key to successful practical applications.

3. Research Program

3.1. Supervised Learning

This part of our research focuses on methods where, given a set of examples of input/output pairs, the goal is to predict the output for a new input, with research on kernel methods, calibration methods, and multi-task learning.

3.2. Unsupervised Learning

We focus here on methods where no output is given and the goal is to find structure of certain known types (e.g., discrete or low-dimensional) in the data, with a focus on matrix factorization, statistical tests, dimension reduction, and semi-supervised learning.

3.3. Parsimony

The concept of parsimony is central to many areas of science. In the context of statistical machine learning, this takes the form of variable or feature selection. The team focuses primarily on structured sparsity, with theoretical and algorithmic contributions.

3.4. Optimization

Optimization in all its forms is central to machine learning, as many of its theoretical frameworks are based at least in part on empirical risk minimization. The team focuses primarily on convex and bandit optimization, with a particular focus on large-scale optimization.

4. Application Domains

4.1. Applications for Machine Learning

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. In collaboration with Institut Curie.
- Text processing: document collection modeling, language models.
- Audio processing: source separation, speech/music processing.
- Neuro-imaging: brain-computer interface (fMRI, EEG, MEG).

5. Highlights of the Year

5.1. Highlights of the Year

5.1.1. Awards

Francis Bach, Lagrange Prize in Continuous Optimization, Society for Industrial and Applied Mathematics 2018

Francis Bach, Best Paper Award, NeurIPS 2018.

Francis Bach included in the report *Highly cited researchers, year 2018*, Clarivate Analytics, 2018

Nicolas Flammarion, PhD thesis award in the *Programme Gaspard Monge*, Fondation Mathématique Jacques Hadamard, 2018.

Adrien Taylor, Tucker Prize (finalist) 2018 (dissertation prize by the Math- ematical Optimization Society for 2015-2017).

Adrien Taylor, IBM/FNRS innovation award 2018 (dissertation prize for original contributions to informatics).

Adrien Taylor, Icteam thesis award 2018 (dissertation award by the icteam institute of UCLouvain, Belgium).

Adrien Taylor, Best paper award 2018 from the journal Optimization Letters for the paper *On the worst-case complexity of the gradient method with exact line search for smooth strongly convex functions*, Etienne De Klerk, François Glineur, Adrien Taylor. journal=.

6. New Software and Platforms

6.1. ProxASAGA

KEYWORD: Optimization

FUNCTIONAL DESCRIPTION: A C++/Python code implementing the methods in the paper "Breaking the Nonsmooth Barrier: A Scalable Parallel Method for Composite Optimization", F. Pedregosa, R. Leblond and S. Lacoste-Julien, Advances in Neural Information Processing Systems (NIPS) 2017. Due to their simplicity and excellent performance, parallel asynchronous variants of stochastic gradient descent have become popular methods to solve a wide range of large-scale optimization problems on multi-core architectures. Yet, despite their practical success, support for nonsmooth objectives is still lacking, making them unsuitable for many problems of interest in machine learning, such as the Lasso, group Lasso or empirical risk minimization with convex constraints. In this work, we propose and analyze ProxASAGA, a fully asynchronous sparse method inspired by SAGA, a variance reduced incremental gradient algorithm. The proposed method is easy to implement and significantly outperforms the state of the art on several nonsmooth, large-scale problems. We prove that our method achieves a theoretical linear speedup with respect to the sequential version under assumptions on the sparsity of gradients and block-separability of the proximal term. Empirical benchmarks on a multi-core architecture illustrate practical speedups of up to 12x on a 20-core machine.

- Contact: Fabian Pedregosa
- URL: https://github.com/fabianp/ProxASAGA

6.2. object-states-action

KEYWORD: Computer vision

FUNCTIONAL DESCRIPTION: Code for the paper Joint Discovery of Object States and Manipulation Actions, ICCV 2017: Many human activities involve object manipulations aiming to modify the object state. Examples of common state changes include full/empty bottle, open/closed door, and attached/detached car wheel. In this work, we seek to automatically discover the states of objects and the associated manipulation actions. Given a set of videos for a particular task, we propose a joint model that learns to identify object states and to localize state-modifying actions. Our model is formulated as a discriminative clustering cost with constraints. We assume a consistent temporal order for the changes in object states and manipulation actions, and introduce new optimization techniques to learn model parameters without additional supervision. We demonstrate successful discovery of seven manipulation actions and corresponding object states on a new dataset of videos depicting real-life object manipulations. We show that our joint formulation results in an improvement of object state discovery by action recognition and vice versa.

- Participants: Jean-Baptiste Alayrac, Josef Sivic, Ivan Laptev and Simon Lacoste-Julien
- Contact: Jean-Baptiste Alayrac
- Publication: Joint Discovery of Object States and Manipulation Actions
- URL: https://github.com/jalayrac/object-states-action

7. New Results

7.1. On the Global Convergence of Gradient Descent for Over-parameterized Models using Optimal Transport

Many tasks in machine learning and signal processing can be solved by minimizing a convex function of a measure. This includes sparse spikes deconvolution or training a neural network with a single hidden layer. For these problems, in [25] we study a simple minimization method: the unknown measure is discretized into a mixture of particles and a continuous-time gradient descent is performed on their weights and positions. This is an idealization of the usual way to train neural networks with a large hidden layer. We show that, when initialized correctly and in the many-particle limit, this gradient flow, although non-convex, converges to global minimizers. The proof involves Wasserstein gradient flows, a by-product of optimal transport theory. Numerical experiments show that this asymptotic behavior is already at play for a reasonable number of particles, even in high dimension.

7.2. Sharp Analysis of Learning with Discrete Losses

In [49], we study a least-squares framework to systematically design learning algorithms for discrete losses, with quantitative characterizations in terms of statistical and computational complexity. In particular we improve existing results by providing explicit dependence on the number of labels for a wide class of losses and faster learning rates in conditions of low-noise. Theoretical results are complemented with experiments on real datasets, showing the effectiveness of the proposed general approach.

7.3. Gossip of Statistical Observations using Orthogonal Polynomials

Consider a network of agents connected by communication links, where each agent holds a real value. The gossip problem consists in estimating the average of the values diffused in the network in a distributed manner. Current techniques for gossiping are designed to deal with worst-case scenarios, which is irrelevant in applications to distributed statistical learning and denoising in sensor networks. In [39], we design second-order gossip methods tailor-made for the case where the real values are i.i.d. samples from the same distribution. In some regular network structures, we are able to prove optimality of our methods, and simulations suggest that they are efficient in a wide range of random networks. Our approach of gossip stems from a new acceleration framework using the family of orthogonal polynomials with respect to the spectral measure of the network graph.

7.4. Marginal Weighted Maximum Log-likelihood for Efficient Learning of Perturb-and-Map models

In [20], We consider the structured-output prediction problem through probabilistic approaches and generalize the “perturb-and-MAP” framework to more challenging weighted Hamming losses, which are crucial in applications. While in principle our approach is a straightforward marginalization, it requires solving many related MAP inference problems. We show that for log-supermodular pairwise models these operations can be performed efficiently using the machinery of dynamic graph cuts. We also propose to use double stochastic gradient descent, both on the data and on the perturbations, for efficient learning. Our framework can naturally take weak supervision (e.g., partial labels) into account. We conduct a set of experiments on medium-scale character recognition and image segmentation, showing the benefits of our algorithms.

7.5. Slice inverse regression with score functions

In [6], we consider non-linear regression problems where we assume that the response depends non-linearly on a linear projection of the covariates. We propose score function extensions to sliced inverse regression problems, both for the first- order and second-order score functions. We show that they provably improve estimation in the population case over the non-sliced versions and we study finite sample estimators and their consistency given the exact score functions. We also propose to learn the score function as well, in two steps, i.e., first learning the score function and then learning the effective dimension reduction space, or directly, by solving a convex optimization problem regularized by the nuclear norm. We illustrate our results on a series of experiments.

7.6. Constant Step Size Stochastic Gradient Descent for Probabilistic Modeling

Stochastic gradient methods enable learning probabilistic models from large amounts of data. While large step-sizes (learning rates) have shown to be best for least-squares (e.g., Gaussian noise) once combined with parameter averaging, these are not leading to convergent algorithms in general. In this paper, we consider generalized linear models, that is, conditional models based on exponential families. In [14], we propose averaging moment parameters instead of natural parameters for constant-step-size stochastic gradient descent. For finite-dimensional models, we show that this can sometimes (and surprisingly) lead to better predictions than the best linear model. For infinite-dimensional models, we show that it always converges to optimal predictions, while averaging natural parameters never does. We illustrate our findings with simulations on synthetic data and classical benchmarks with many observations.

7.7. Nonlinear Acceleration of Momentum and Primal-Dual Algorithms

In [40], We describe a convergence acceleration scheme for multistep optimization algorithms. The extrapolated solution is written as a nonlinear average of the iterates produced by the original optimization algorithm. Our scheme does not need the underlying fixed-point operator to be symmetric, hence handles e.g. algorithms with momentum terms such as Nesterov's accelerated method, or primal-dual methods. The weights are computed via a simple linear system and we analyze performance in both online and offline modes. We use Crouzeix's conjecture to show that acceleration performance is controlled by the solution of a Chebyshev problem on the numerical range of a non-symmetric operator modelling the behavior of iterates near the optimum. Numerical experiments are detailed on image processing problems, logistic regression and neural network training for CIFAR10 and ImageNet.

7.8. Nonlinear Acceleration of Deep Neural Networks

Regularized nonlinear acceleration (RNA) is a generic extrapolation scheme for optimization methods, with marginal computational overhead. It aims to improve convergence using only the iterates of simple iterative algorithms. However, so far its application to optimization was theoretically limited to gradient descent and other single-step algorithms. Here, we adapt RNA to a much broader setting including stochastic gradient with momentum and Nesterov's fast gradient. In [54], we use it to train deep neural networks, and empirically observe that extrapolated networks are more accurate, especially in the early iterations. A straightforward application of our algorithm when training ResNet-152 on ImageNet produces a top-1 test error of 20.88, improving by 0.8 the reference classification pipeline. Furthermore, the code runs offline in this case, so it never negatively affects performance.

7.9. Nonlinear Acceleration of CNNs

The Regularized Nonlinear Acceleration (RNA) algorithm is an acceleration method capable of improving the rate of convergence of many optimization schemes such as gradient descend, SAGA or SVRG. Until now, its analysis is limited to convex problems, but empirical observations shows that RNA may be extended to wider settings. In [36], we investigate further the benefits of RNA when applied to neural networks, in particular for the task of image recognition on CIFAR10 and ImageNet. With very few modifications of exiting frameworks, RNA improves slightly the optimization process of CNNs, after training.

7.10. Robust Seriation and Applications To Cancer Genomics

The seriation problem seeks to reorder a set of elements given pairwise similarity information, so that elements with higher similarity are closer in the resulting sequence. When a global ordering consistent with the similarity information exists, an exact spectral solution recovers it in the noiseless case and seriation is equivalent to the combinatorial 2-SUM problem over permutations, for which several relaxations have been derived. However, in applications such as DNA assembly, similarity values are often heavily corrupted, and the solution of 2-SUM may no longer yield an approximate serial structure on the elements. In [52], we introduce the robust seriation problem and show that it is equivalent to a modified 2-SUM problem for a class of similarity matrices modeling those observed in DNA assembly. We explore several relaxations of this modified 2-SUM problem and compare them empirically on both synthetic matrices and real DNA data. We then introduce the problem of seriation with duplications, which is a generalization of Seriation motivated by applications to cancer genome reconstruction. We propose an algorithm involving robust seriation to solve it, and present preliminary results on synthetic data sets.

7.11. Reconstructing Latent Orderings by Spectral Clustering

Spectral clustering uses a graph Laplacian spectral embedding to enhance the cluster structure of some data sets. When the embedding is one dimensional, it can be used to sort the items (spectral ordering). A number of empirical results also suggests that a multidimensional Laplacian embedding enhances the latent ordering of the data, if any. This also extends to circular orderings, a case where unidimensional embeddings fail. In [51], we tackle the task of retrieving linear and circular orderings in a unifying framework, and show how a latent ordering on the data translates into a filamentary structure on the Laplacian embedding. We propose a method to recover it, illustrated with numerical experiments on synthetic data and real DNA sequencing data.

7.12. Lyapunov Functions for First-Order Methods: Tight Automated Convergence Guarantees

In [21], we present a novel way of generating Lyapunov functions for proving linear convergence rates of first-order optimization methods. Our approach provably obtains the fastest linear convergence rate that can be verified by a quadratic Lyapunov function (with given states), and only relies on solving a small-sized semidefinite program. Our approach combines the advantages of performance estimation problems and integral quadratic constraints, and relies on convex interpolation.

7.13. Efficient First-order Methods for Convex Minimization: a Constructive Approach

In [44], we describe a novel constructive technique for devising efficient first-order methods for a wide range of large-scale convex minimization settings, including smooth, non-smooth, and strongly convex minimization. The design technique takes a method performing a series of subspace-searches and constructs a family of methods that share the same worst-case guarantees as the original method, and includes a fixed-step first-order method. We show that this technique yields optimal methods in the smooth and non-smooth cases and derive new methods for these cases, including methods that forego knowledge of the problem parameters, at the cost of a one-dimensional line search per iteration. In the strongly convex case, we show how numerical tools can be used to perform the construction, and show that resulting method offers an improved convergence rate compared to Nesterov's celebrated fast gradient method.

7.14. Operator Splitting Performance Estimation: Tight contraction factors and optimal parameter selection

In [53], we propose a methodology for studying the performance of common splitting methods through semidefinite programming. We prove tightness of the methodology and demonstrate its value by presenting two applications of it. First, we use the methodology as a tool for computer-assisted proofs to prove tight analytical contraction factors for Douglas–Rachford splitting that are likely too complicated for a human to find bare-handed. Second, we use the methodology as an algorithmic tool to computationally select the optimal splitting method parameters by solving a series of semidefinite programs.

7.15. Finite-sample Analysis of M-estimators using Self-concordance

In [50], we demonstrate how *self-concordance* of the loss allows to obtain asymptotically optimal rates for M-estimators in finite-sample regimes. We consider two classes of losses: (i) self-concordant losses, i.e., whose third derivative is uniformly bounded with the $3/2$ power of the second; (ii) *pseudo* self-concordant losses, for which the power is removed. These classes contain some losses arising in generalized linear models, including the logistic loss; in addition, the second class includes some common pseudo-Huber losses. Our results consist in establishing the *critical sample size* sufficient to reach the asymptotically optimal excess risk in both cases. Denoting d the parameter dimension, and d_e the effective dimension taking into account possible model misspecification, we find the critical sample size to be $O(d_e \cdot d)$ for the first class of losses, and $O(\rho \cdot d_e \cdot d)$ for the second class, where ρ is the problem-dependent parameter that characterizes the risk curvature at the best predictor θ_*. In contrast to the existing results, we only impose *local* assumptions on the data distribution, assuming that the calibrated design, i.e., the design scaled with the square root of the second derivative of the loss, is subgaussian at the best predictor. Moreover, we obtain the improved bounds on the critical sample size, scaling *near-linearly* in $\max(d_e, d)$, under the extra assumption that the calibrated design is subgaussian in the Dikin ellipsoid of θ_*. Motivated by these findings, we construct canonically self-concordant analogues of the Huber and logistic losses with improved statistical properties. Finally, we extend some of the above results to ℓ_1-penalized M-estimators in high-dimensional setups.

7.16. Uniform regret bounds over R^d for the sequential linear regression problem with the square loss

In [45] we consider the setting of online linear regression for arbitrary deterministic sequences, with the square loss. We are interested in obtaining regret bounds that hold uniformly over all vectors R^d. When the feature sequence is known at the beginning of the game, they provided closed-form regret bounds of $2dB^2 \ln T + O(1)$, where T is the number of rounds and B is a bound on the observations. Instead, we derive bounds with an optimal constant of 1 in front of the $dB^2 \ln T$ term. In the case of sequentially revealed features, we also derive an asymptotic regret bound of $dB^2 \ln T$ for any individual sequence of features and bounded observations. All our algorithms are variants of the online nonlinear ridge regression forecaster, either with a data-dependent regularization or with almost no regularization.

7.17. Efficient online algorithms for fast-rate regret bounds under sparsity.

In [46] we consider the problem of online convex optimization in two different settings: arbitrary and i.i.d. sequence of convex loss functions. In both settings, we provide efficient algorithms whose cumulative excess risks are controlled with fast-rate sparse bounds. First, the excess risks bounds depend on the sparsity of the objective rather than on the dimension of the parameters space. Second, their rates are faster than the slow-rate $1/\sqrt{T}$

7.18. Exponential convergence of testing error for stochastic gradient methods

In [32], we consider binary classification problems with positive definite kernels and square loss, and study the convergence rates of stochastic gradient methods. We show that while the excess testing loss (squared loss) converges slowly to zero as the number of observations (and thus iterations) goes to infinity, the testing error (classification error) converges exponentially fast if low-noise conditions are assumed.

7.19. Statistical Optimality of Stochastic Gradient Descent on Hard Learning Problems through Multiple Passes

In [33], we consider stochastic gradient descent (SGD) for least-squares regression with potentially several passes over the data. While several passes have been widely reported to perform practically better in terms of predictive performance on unseen data, the existing theoretical analysis of SGD suggests that a single pass is statistically optimal. While this is true for low-dimensional easy problems, we show that for hard problems, multiple passes lead to statistically optimal predictions while single pass does not; we also show that in these hard models, the optimal number of passes over the data increases with sample size. In order to define the notion of hardness and show that our predictive performances are optimal, we consider potentially infinite-dimensional models and notions typically associated to kernel methods, namely, the decay of eigenvalues of the covariance matrix of the features and the complexity of the optimal predictor as measured through the covariance matrix. We illustrate our results on synthetic experiments with non-linear kernel methods and on a classical benchmark with a linear model.

7.20. Central Limit Theorem for stationary Fleming–Viot particle systems in finite spaces

In [11], we consider the Fleming–Viot particle system associated with a continuous-time Markov chain in a finite space. Assuming irreducibility, it is known that the particle system possesses a unique stationary distribution, under which its empirical measure converges to the quasistationary distribution of the Markov chain. We complement this Law of Large Numbers with a Central Limit Theorem. Our proof essentially relies on elementary computations on the infinitesimal generator of the Fleming–Viot particle system, and involves the so-called π-return process in the expression of the asymptotic variance. Our work can be seen as an infinite-time version, in the setting of finite space Markov chains, of results by Del Moral and Miclo [ESAIM: Probab. Statist., 2003] and Cérou, Delyon, Guyader and Rousset [arXiv:1611.00515, arXiv:1709.06771].

7.21. SeaRNN: Improved RNN training through Global-Local Losses

In [16], we propose SEARNN, a novel training algorithm for recurrent neural networks (RNNs) inspired by the “learning to search” (L2S) approach to structured prediction. RNNs have been widely successful in structured prediction applications such as machine translation or parsing, and are commonly trained using maximum likelihood estimation (MLE). Unfortunately, this training loss is not always an appropriate surrogate for the test error: by only maximizing the ground truth probability, it fails to exploit the wealth of information offered by structured losses. Further, it introduces discrepancies between training and predicting (such as exposure bias) that may hurt test performance. Instead, SEARNN leverages test-alike search space exploration to introduce global-local losses that are closer to the test error. We first demonstrate improved performance over MLE on two different tasks: OCR and spelling correction. Then, we propose a subsampling strategy to enable SEARNN to scale to large vocabulary sizes. This allows us to validate the benefits of our approach on a machine translation task.

7.22. Improved asynchronous parallel optimization analysis for stochastic incremental methods

As datasets continue to increase in size and multi-core computer architectures are developed, asynchronous parallel optimization algorithms become more and more essential to the field of Machine Learning. Unfortunately, conducting the theoretical analysis of asynchronous methods is difficult, notably due to the introduction of delay and inconsistency in inherently sequential algorithms. Handling these issues often requires resorting to simplifying but unrealistic assumptions. Through a novel perspective, in [10] we revisit and clarify a subtle but important technical issue present in a large fraction of the recent convergence rate proofs for asynchronous parallel optimization algorithms, and propose a simplification of the recently introduced "perturbed iterate" framework that resolves it. We demonstrate the usefulness of our new framework by analyzing three distinct asynchronous parallel incremental optimization algorithms: Hogwild (asynchronous SGD), KROMAGNON (asynchronous SVRG) and ASAGA, a novel asynchronous parallel version of the incremental gradient algorithm SAGA that enjoys fast linear convergence rates. We are able to both remove problematic assumptions and obtain better theoretical results. Notably, we prove that ASAGA and KROMAGNON can obtain a theoretical linear speedup on multi-core systems even without sparsity assumptions. We present results of an implementation on a 40-core architecture illustrating the practical speedups as well as the hardware overhead. Finally, we investigate the overlap constant, an ill-understood but central quantity for the theoretical analysis of asynchronous parallel algorithms. We find that it encompasses much more complexity than suggested in previous work, and often is order-of-magnitude bigger than traditionally thought.

7.23. Asynchronous optimisation for Machine Learning

The impressive breakthroughs of the last two decades in the field of machine learning can be in large part attributed to the explosion of computing power and available data. These two limiting factors have been replaced by a new bottleneck: algorithms. The focus of this thesis [3] is thus on introducing novel methods that can take advantage of high data quantity and computing power. We present two independent contributions.

First, we develop and analyze novel fast optimization algorithms which take advantage of the advances in parallel computing architecture and can handle vast amounts of data. We introduce a new framework of analysis for asynchronous parallel incremental algorithms, which enable correct and simple proofs. We then demonstrate its usefulness by performing the convergence analysis for several methods, including two novel algorithms.

Asaga is a sparse asynchronous parallel variant of the variance-reduced algorithm Saga which enjoys fast linear convergence rates on smooth and strongly convex objectives. We prove that it can be linearly faster than its sequential counterpart, even without sparsity assump- tions.

ProxAsaga is an extension of Asaga to the more general setting where the regularizer can be non-smooth. We prove that it can also achieve a linear speedup. We provide extensive experiments comparing our new algorithms to the current state-of-art.

Second, we introduce new methods for complex struc- tured prediction tasks. We focus on recurrent neural net- works (RNNs), whose traditional training algorithm for RNNs – based on maximum likelihood estimation (MLE) – suffers from several issues. The associated surrogate training loss notably ignores the information contained in structured losses and introduces discrepancies between train and test times that may hurt performance.

To alleviate these problems, we propose SeaRNN, a novel training algorithm for RNNs inspired by the "learning to search" approach to structured prediction.SeaRNN leverages test-alike search space exploration to introduce global-local losses that are closer to the test error than the MLE objective.

We demonstrate improved performance over MLE on three challenging tasks, and provide several subsampling strategies to enable SeaRNN to scale to large-scale tasks, such as machine translation. Finally, after contrasting the behavior of SeaRNN models to MLE models, we conduct an in-depth comparison of our new approach to the related work.

7.24. M^*-Regularized Dictionary Learning

In [38], we derive a performance measure for dictionaries in compressed sensing, based on the M^* of the corresponding norm. We use this measure to regularize dictionary learning algorithms and study the performance of our methods on both compression and inpainting experiments.

7.25. Optimal Algorithms for Non-Smooth Distributed Optimization in Networks

In [35], we consider the distributed optimization of non-smooth convex functions using a network of computing units. We investigate this problem under two regularity assumptions: (1) the Lipschitz continuity of the global objective function, and (2) the Lipschitz continuity of local individual functions. Under the local regularity assumption, we provide the first optimal first-order decentralized algorithm called multi-step primal-dual (MSPD) and its corresponding optimal convergence rate. A notable aspect of this result is that, for non-smooth functions, while the dominant term of the error is in $O(1/\sqrt{t})$, the structure of the communication network only impacts a second-order term in $O(1/t)$, where t is time. In other words, the error due to limits in communication resources decreases at a fast rate even in the case of non-strongly-convex objective functions. Under the global regularity assumption, we provide a simple yet efficient algorithm called distributed randomized smoothing (DRS) based on a local smoothing of the objective function, and show that DRS is within a $d^{1/4}$ multiplicative factor of the optimal convergence rate, where d is the underlying dimension.

7.26. Relating Leverage Scores and Density using Regularized Christoffel Functions

Statistical leverage scores emerged as a fundamental tool for matrix sketching and column sampling with applications to low rank approximation, regression, random feature learning and quadrature. Yet, the very nature of this quantity is barely understood. Borrowing ideas from the orthogonal polynomial literature, we introduce in [31] the regularized Christoffel function associated to a positive definite kernel. This uncovers a variational formulation for leverage scores for kernel methods and allows to elucidate their relationships with the chosen kernel as well as population density. Our main result quantitatively describes a decreasing relation between leverage score and population density for a broad class of kernels on Euclidean spaces. Numerical simulations support our findings.

7.27. Averaging Stochastic Gradient Descent on Riemannian Manifolds

In [37] we consider the minimization of a function defined on a Riemannian manifold M accessible only through unbiased estimates of its gradients. We develop a geometric framework to transform a sequence of slowly converging iterates generated from stochastic gradient descent (SGD) on M to an averaged iterate sequence with a robust and fast $O(1/n)$ convergence rate. We then present an application of our framework to geodesically-strongly-convex (and possibly Euclidean non-convex) problems. Finally, we demonstrate how these ideas apply to the case of streaming k-PCA, where we show how to accelerate the slow rate of the randomized power method (without requiring knowledge of the eigengap) into a robust algorithm achieving the optimal rate of convergence.

7.28. Localized Structured Prediction

Key to structured prediction is exploiting the problem structure to simplify the learning process. A major challenge arises when data exhibit a local structure (e.g., are made by "parts") that can be leveraged to better approximate the relation between (parts of) the input and (parts of) the output. Recent literature on signal processing, and in particular computer vision, has shown that capturing these aspects is indeed essential to achieve state-of-the-art performance. While such algorithms are typically derived on a case-by-case basis, in [42] we propose the first theoretical framework to deal with part-based data from a general perspective. We derive a novel approach to deal with these problems and study its generalization properties within the setting of statistical learning theory. Our analysis is novel in that it explicitly quantifies the benefits of leveraging the part-based structure of the problem with respect to the learning rates of the proposed estimator.

7.29. Optimal rates for spectral algorithms with least-squares regression over Hilbert spaces

In [12], we study regression problems over a separable Hilbert space with the square loss, covering non-parametric regression over a reproducing kernel Hilbert space. We investigate a class of spectral-regularized algorithms, including ridge regression, principal component analysis, and gradient methods. We prove optimal, high-probability convergence results in terms of variants of norms for the studied algorithms, considering a capacity assumption on the hypothesis space and a general source condition on the target function. Consequently, we obtain almost sure convergence results with optimal rates. Our results improve and generalize previous results, filling a theoretical gap for the non-attainable cases.

7.30. Differential Properties of Sinkhorn Approximation for Learning with Wasserstein Distance

Applications of optimal transport have recently gained remarkable attention thanks to the computational advantages of entropic regularization. However, in most situations the Sinkhorn approximation of the Wasserstein distance is replaced by a regularized version that is less accurate but easy to differentiate. In [17] we characterize the differential properties of the original Sinkhorn distance, proving that it enjoys the same smoothness as its regularized version and we explicitly provide an efficient algorithm to compute its gradient. We show that this result benefits both theory and applications: on one hand, high order smoothness confers statistical guarantees to learning with Wasserstein approximations. On the other hand, the gradient formula allows us to efficiently solve learning and optimization problems in practice. Promising preliminary experiments complement our analysis.

7.31. Learning with SGD and Random Features

Sketching and stochastic gradient methods are arguably the most common techniques to derive efficient large scale learning algorithms. In [15], we investigate their application in the context of nonparametric statistical learning. More precisely, we study the estimator defined by stochastic gradient with mini batches and random features. The latter can be seen as form of nonlinear sketching and used to define approximate kernel methods. The considered estimator is not explicitly penalized/constrained and regularization is implicit. Indeed, our study highlights how different parameters, such as number of features, iterations, step-size and mini-batch size control the learning properties of the solutions. We do this by deriving optimal finite sample bounds, under standard assumptions. The obtained results are corroborated and illustrated by numerical experiments.

7.32. Manifold Structured Prediction

Structured prediction provides a general framework to deal with supervised problems where the outputs have semantically rich structure. While classical approaches consider finite, albeit potentially huge, output spaces, in [19] we discuss how structured prediction can be extended to a continuous scenario. Specifically, we study a structured prediction approach to manifold valued regression. We characterize a class of problems for which the considered approach is statistically consistent and study how geometric optimization can be used to compute the corresponding estimator. Promising experimental results on both simulated and real data complete our study.

7.33. On Fast Leverage Score Sampling and Optimal Learning

Leverage score sampling provides an appealing way to perform approximate computations for large matrices. Indeed, it allows to derive faithful approximations with a complexity adapted to the problem at hand. Yet, performing leverage scores sampling is a challenge in its own right requiring further approximations. In [18], we study the problem of leverage score sampling for positive definite matrices defined by a kernel. Our contribution is twofold. First we provide a novel algorithm for leverage score sampling and second, we exploit the proposed method in statistical learning by deriving a novel solver for kernel ridge regression. Our main technical contribution is showing that the proposed algorithms are currently the most efficient and accurate for these problems.

7.34. Accelerated Decentralized Optimization with Local Updates for Smooth and Strongly Convex Objectives

In [47], we study the problem of minimizing a sum of smooth and strongly convex functions split over the nodes of a network in a decentralized fashion. We propose a decentralized accelerated algorithm that only requires local synchrony. Its rate depends on the condition number κ of the local functions as well as the network topology and delays. Under mild assumptions on the topology of the graph, our algorithm takes a time $O((\tau_{\max} + \Delta_{\max})\sqrt{\kappa/\gamma}\ln(\epsilon^{-1}))$ to reach a precision ϵ where γ is the spectral gap of the graph, $\tau_{\max}$ the maximum communication delay and $\Delta_{\max}$ the maximum computation time. Therefore, it matches the rate of SSDA, which is optimal when $\tau_{\max} = \Omega\left(\Delta_{\max}\right)$. Applying our algorithm to quadratic local functions leads to an accelerated randomized gossip algorithm of rate $O(\sqrt{\theta_{\text{gossip}}/n})$ where θ_{gossip} is the rate of the standard randomized gossip. To the best of our knowledge, it is the first asynchronous gossip algorithm with a provably improved rate of convergence of the second moment of the error. We illustrate these results with experiments in idealized settings.

8. Bilateral Contracts and Grants with Industry

8.1. Bilateral Contracts with Industry

Microsoft Research: “Structured Large-Scale Machine Learning”. Machine learning is now ubiquitous in industry, science, engineering, and personal life. While early successes were obtained by applying off-the-shelf techniques, there are two main challenges faced by machine learning in the “big data” era: structure and scale. The project proposes to explore three axes, from theoretical, algorithmic and practical perspectives: (1) large-scale convex optimization, (2) large-scale combinatorial optimization and (3) sequential decision making for structured data. The project involves two Inria sites (Paris and Grenoble) and four MSR sites (Cambridge, New England, Redmond, New York). Project website: http://www.msr-inria.fr/projects/structured-large-scale-machine-learning/.

8.2. Bilateral Grants with Industry

- Alexandre d’Aspremont, Francis Bach, Martin Jaggi (EPFL): Google Focused award.
- Francis Bach: Gift from Facebook AI Research.
- Alexandre d’Aspremont: AXA, "mécénat scientifique, chaire Havas-Dauphine", machine learning.

9. Partnerships and Cooperations

9.1. National Initiatives

Alexandre d’Aspremont: IRIS, PSL “Science des données, données de la science”.

9.2. European Initiatives

- **ITN Spartan**
  Title: Sparse Representations and Compressed Sensing Training Network
  Type: FP7
  Instrument: Initial Training Network
  Duration: October 2014 to October 2018
  Coordinator: Mark Plumbley (University of Surrey)
  Inria contact: Francis Bach
  Abstract: The SpaRTaN Initial Training Network will train a new generation of interdisciplinary researchers in sparse representations and compressed sensing, contributing to Europe’s leading role

in scientific innovation. By bringing together leading academic and industry groups with expertise in sparse representations, compressed sensing, machine learning and optimisation, and with an interest in applications such as hyperspectral imaging, audio signal processing and video analytics, this project will create an interdisciplinary, trans-national and inter-sectorial training network to enhance mobility and training of researchers in this area. SpaRTaN is funded under the FP7-PEOPLE-2013-ITN call and is part of the Marie Curie Actions — Initial Training Networks (ITN) funding scheme: Project number - 607290

- **ITN Macsenet**
  Title: Machine Sensing Training Network
  Type: H2020
  Instrument: Initial Training Network
  Duration: January 2015 - January 2019
  Coordinator: Mark Plumbley (University of Surrey)
  Inria contact: Francis Bach
  Abstract: The aim of this Innovative Training Network is to train a new generation of creative, entrepreneurial and innovative early stage researchers (ESRs) in the research area of measurement and estimation of signals using knowledge or data about the underlying structure. We will develop new robust and efficient Machine Sensing theory and algorithms, together methods for a wide range of signals, including: advanced brain imaging; inverse imaging problems; audio and music signals; and non-traditional signals such as signals on graphs. We will apply these methods to real-world problems, through work with non-Academic partners, and disseminate the results of this research to a wide range of academic and non-academic audiences, including through publications, data, software and public engagement events. MacSeNet is funded under the H2020-MSCA-ITN-2014 call and is part of the Marie Sklodowska- Curie Actions — Innovative Training Networks (ITN) funding scheme.
- **ERC Sequoia** Title: Robust algorithms for learning from modern data
  Programm: H2020
  Type: ERC
  Duration: 2017-2022
  Coordinator: Inria
  Inria contact: Francis Bach
  Abstract: 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 project 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, bioin- formatics, 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.

9.3. International Initiatives

9.3.1. BigFOKS2

Title: Learning from Big Data: First-Order methods for Kernels and Submodular functions

International Partner (Institution - Laboratory - Researcher):

IISc Bangalore (India) - Computer Science Department - Chiranjib Bhattacharyya

Start year: 2016

See also: mllab.csa.iisc.ernet.in/indo-french.html

Recent advances in sensor technologies have resulted in large amounts of data being generated in a wide array of scientific disciplines. Deriving models from such large datasets, often known as “Big Data”, is one of the important challenges facing many engineering and scientific disciplines. In this proposal we investigate the problem of learning supervised models from Big Data, which has immediate applications in Computational Biology, Computer vision, Natural language processing, Web, E-commerce, etc., where specific structure is often present and hard to take into account with current algorithms. Our focus will be on the algorithmic aspects. Often supervised learning problems can be cast as convex programs. The goal of this proposal will be to derive first-order methods which can be effective for solving such convex programs arising in the Big-Data setting. Keeping this broad goal in mind we investigate two foundational problems which are not well addressed in existing literature. The first problem investigates Stochastic Gradient Descent Algorithms in the context of First-order methods for designing algorithms for Kernel based prediction functions on Large Datasets. The second problem involves solving discrete optimization problems arising in Submodular formulations in Machine Learning, for which first-order methods have not reached the level of speed required for practical applications (notably in computer vision).

9.4. International Research Visitors

- Vijaya Bollapragada from Northwestern University, Chicago, IL, United States, Apr - Jul 2018.
- Aaron De Fazio from Facebook Research NY, New York, United States, Feb 2018.
- Gauthier Gidel from University of Montreal - MILA, Montreal, Canada, Jan 2018.
- Sharan Vaswani from University of British Columbia, Vancouver, Canada, Apr - Jul 2018
- Simon Lacoste-Julien from University of Montreal - MILA, Montreal, Canada, Aug 2018.

10. Dissemination

10.1. Promoting Scientific Activities

10.1.1. Scientific Events Organisation

10.1.1.1. General Chair, Scientific Chair

F. Bach: General Chair of ICML 2018 (Stockholm)

10.1.1.2. Member of the Organizing Committees

Adrian Taylor, Session Organizer: *Computer-assisted analyses of optimization algorithms I & II*, International Symposium on Mathematical Programming, July 2018.

F. Bach: Co-organization of the workshop “Horizon Maths 2018 : Intelligence Artificielle”, November 23, 2018

10.1.2. Scientific Events Selection

10.1.2.1. Chair of Conference Program Committees

F. Bach: Program Chair of the Journées de Statistiques (Saclay)

10.1.2.2. Reviewer

Conference on Learning Theory (COLT 2018): Pierre Gaillard, Alessandro Rudi

Symposium on Discrete Algorithms (SODA 2019): Adrien Taylor,

Neural Information Processing Systems (NIPS 2018): Pierre Gaillard, Alessandro Rudi

Conference on Learning Theory (COLT 2018): Pierre Gaillard, Alessandro Rudi, Adrien Taylor

Symposium on Discrete Algorithms (SODA 2019): Adrien Taylor

International Conference of Machine Learning (ICML 2018): Pierre Gaillard, Alessandro Rudi

10.1.3. Journal

10.1.3.1. Member of the Editorial Boards

F. Bach: Journal of Machine Learning Research, co-editor-in-chief

F. Bach: Information and Inference, Associate Editor.

F. Bach: Electronic Journal of Statistics, Associate Editor.

F. Bach: Mathematical Programming, Associate Editor.

F. Bach: Foundations of Computational Mathematics, Associate Editor.

A. d'Aspremont: SIAM Journal on Optimization, Associate editor

A. d'Aspremont: SIAM Journal on the Mathematics of Data Science, Associate Editor

A. d'Aspremont: Mathematical Programming, Associate Editor

10.1.3.2. Reviewer - Reviewing Activities

SIAM Journal on Optimization: Adrien Taylor

Mathematical Programming: Adrien Taylor

Journal of Optimization Theory and Algorithms: Adrien Taylor

Journal of Machine Learning Research: Pierre Gaillard, Alessandro Rudi

Applied Computational Harmonic Analysis: Alessandro Rudi

10.1.4. Invited Talks

F. Bach, Trends in Optimization Seminar, University of Washington, November 2018.

Pierre Gaillard. *Distributed averaging of observations in a graph: the gossip problem*. MNL Conference, Paris, November 2018.

Adrien Taylor, *Analysis and design of first-order methods via semidefinite programming*, Seminaire Parisien dOptimisation (SPO), Paris (France), November 2018.

F. Bach, Frontier Research and Artificial Intelligence, European Research Council, Brussels, October 2018.

F. Bach, IDSS Distinguished Speaker Seminar, MIT, October 2018.

F. Bach, Mathematical Institute Colloquium, Oxford, October 2018.

Adrien Taylor, *Convex Interpolation and Performance Estimation of First- order Methods* for Convex Optimization, IBM/FNRS innovation award, Brussels (Belgium), October 2018.

F. Bach, Workshop on Structural Inference in High-Dimensional Models, Moscow, September 2018.

F. Bach, Symposium on Mathematical Programming (ISMP), Bordeaux, plenary talk, July 2018.

Alexandre d'Aspremont, *Sharpness, Restart and Compressed Sensing Performance*, ISMP 2018, Bordeaux, July 2018.

Alessandro Rudi, *FALKON: An optimal method for large scale learning with statistical guarantees*, ISMP 2018, Bordeaux, July 2018.

Adrien Taylor, *Computer-assisted Lyapunov-based worst-case analyses of first- order methods*, International Symposium on Mathematical Programming, Bordeaux (France), July 2018.

F. Bach, SIAM Conference on Imaging Science, Bologna, Italy, invited talk, June 2018.

Pierre Gaillard. *Online prediction of arbitrary time-series with application to electricity consumption*. Conference on nonstationarity. Cergy Pontoise University. June 2018.

Adrien Taylor, *Convex Interpolation and Performance Estimation of First-order Methods for Convex Optimization*, International Symposium on Mathematical Programming: Tucker prize finalist, Bordeaux (France), July 2018.

Alexandre d'Aspremont, *An approximate Shapley-Folkman Theorem*, Isaac Newton Institute, Cambridge, June 2018.

F. Bach,Workshop on Future challenges in statistical scalability, Newton Institute, Cambridge, UK, June 2018.

Adrien Taylor, *Automated design of first-order optimization methods*, Operation Research Seminar, UCLouvain, Louvain-la-Neuve (Belgium), May 2018.

Adrien Taylor, *Automated design of first-order optimization methods*, LCCC Control Seminar, Lund University, Lund (Sweden), May 2018.

Pierre Gaillard. *Distributed learning with orthogonal polynomials*. Inria DGA meetup. May 2018.

F. Bach, Workshop on Optimisation and Machine Learning in Economics, London, March 2018.

Pierre Gaillard. *An overview of Artificial Intelligence*. Hackaton. PSL University. March 2018.

Alexandre d'Aspremont, *Regularized Nonlinear Acceleration*, US and Mexico Workshop on Optimization and its Applications, Jan 2018.

Alessandro Rudi, *Learning with Random Features*, Isaac Newton Institute, Cambridge, Jan 2018.

Pierre Gaillard. *Online nonparametric regression with adversarial data.* Smile seminar. Paris. Jan 2018.

10.2. Teaching - Supervision - Juries

10.2.1. Teaching

F. Bach (together with N. Chopin), *Graphical models*, 30h, Master M2 (MVA), ENS Cachan, France.

F. Bach, *Optimisation et apprentissage statistique*, 20h, Master M2 (Mathématiques de l'aléatoire), Université Paris-Sud, France.

Alexandre d'Aspremont, *Optimisation Combinatoire et Convexe*, avec Zhentao Li, (2015-Present) cours magistraux 30h, Master M1, ENS Paris.

Alexandre d'Aspremont, *Optimisation convexe: modélisation, algorithmes et applications* cours magistraux 21h (2011-Present), Master M2 MVA, ENS PS.

F. Bach and P. Gaillard, *Apprentissage statistique*, 35h, Master M1, Ecole Normale Supérieure, France.

P. Gaillard (together with V. Perchet), *Prediction of individual sequences*, 21h, Master M2 MVA, ENS Cachan, France.

Gregoire Mialon, Python for Machine Learning, 21h, M2 MASH, Dauphine-ENS-PSL, Paris.

10.2.2. Supervision

Anaël Bonneton, PhD defended on July 2018, co-advised by Francis Bach, located in Agence nationale de la sécurité des systèmes d'information (ANSSI).

Damien Scieur, PhD defended on September 2018. *Sur l'accélération des méthodes d'optimisation*, supervised by Alexandre d'Aspremont and Francis Bach.

Jean-Baptiste Alayrac, PhD defended on September 2018, *Structured Learning from Videos and Language*, supervised by Simon Lacoste-Julien, Josef Sivic and Ivan Laptev.

Antoine Recanati, PhD. defended on November 2018. *Application du problème de sériation au séquençage de l'ADN et autres relaxations convexes appliquées en bioinformatique*, supervised by Alexandre d'Aspremont.

Rémi Leblond, PhD defended on November 2018, *Asynchronous Optimization for Machine Learning*, supervised by Simon Lacoste-Julien.

Mathieu Barre, PhD in progress *Méthodes d'extrapolation, au-delà de la convexité*, supervised by Alexandre d'Aspremont.

Grégoire Mialon, PhD in progress *Algorithmes d'optimisation, méthodes de régularisation et architectures pour les réseaux de neurones profonds dans un contexte où les données labellisées sont rares*, supervised by Alexandre d'Aspremont.

Radu-Alexandru Dragomir, PhD in progress *Non-Euclidean first-order methods*, supervised by Alexandre d'Aspremont and Jérôme Bolte.

Thomas Kerdreux, PhD in progress *Optimisation and machine learning*, supervised by Alexandre d'Aspremont.

Margaux Brégère, PhD in progress started September 2017, supervised by Pierre Gaillard, Gilles Stoltz and Yannig Goude (EDF R&D).

Raphaël Berthier, PhD in progress started September 2017, supervised by Francis Bach and Pierre Gaillard.

Loucas Pillaud-Vivien, PhD in progress, supervised by Francis Bach and Alessandro Rudi.

Alex Nowak, PhD in progress, supervised by Francis Bach and Alessandro Rudi.

Ulysse Marteau Ferey, PhD in progress, supervised by Francis Bach and Alessandro Rudi.

Dmitry Babichev, PhD in progress, started is September 2015, co-advised by Francis Bach and Anatoly Judistky (Univ. Grenoble).

Tatiana Shpakova, PhD in progress, started September 2015, advised by Francis Bach.

10.2.3. Juries

Alexandre d'Aspremont, Habilitation à diriger des recherches. Thomas Bruls, Genoscope, Université d'Evry.

10.3. Popularization

10.3.1. Creation of media or tools for science outreach

Design and implementation of a demonstration for the permanent exhibit at Palais de la Découverte: "L'apprenti illustrateur" (J.-B. Alayrac, F. Bach)

11. Bibliography

Publications of the year

Doctoral Dissertations and Habilitation Theses

[1] J.-B. ALAYRAC. *Structured Learning from Videos and Language*, Ecole normale supérieure - ENS PARIS, September 2018, https://hal.inria.fr/tel-01885412

[2] A. BEAUGNON. *Expert-in-the-Loop Supervised Learning for Computer Security Detection Systems*, PSL Research University, June 2018, https://hal.archives-ouvertes.fr/tel-01888971

[3] R. LEBLOND. *Asynchronous Optimization for Machine Learning*, Ecole Normale Superieure de Paris - ENS Paris, November 2018, https://hal.inria.fr/tel-01950576

[4] A. RECANATI. *Relaxations of the Seriation problem and applications to de novo genome assembly*, PSL Research University, November 2018, https://hal.archives-ouvertes.fr/tel-01984368

[5] D. SCIEUR. *Acceleration in Optimization*, PSL Research University, September 2018, https://hal.archives-ouvertes.fr/tel-01887163

Articles in International Peer-Reviewed Journals

[6] D. BABICHEV, F. BACH. *Slice inverse regression with score functions*, in "Electronic journal of statistics ", May 2018, vol. Volume 12, Number 1 (2018), pp. 1507-1543 [*DOI :* 10.1214/18-EJS1428], https://hal.inria.fr/hal-01388498

[7] F. BACH. *Submodular Functions: from Discrete to Continous Domains*, in "Mathematical Programming, Series A", 2018, https://arxiv.org/abs/1511.00394 , https://hal.archives-ouvertes.fr/hal-01222319

[8] A. D'ASPREMONT, C. GUZMAN, M. JAGGI. *Optimal Affine-Invariant Smooth Minimization Algorithms*, in "SIAM Journal on Optimization", July 2018, vol. 28, n^{o} 3, pp. 2384 - 2405 [*DOI :* 10.1137/17M1116842], https://hal.archives-ouvertes.fr/hal-01927392

[9] D. GARREAU, S. ARLOT. *Consistent change-point detection with kernels*, in "Electronic journal of statistics ", December 2018, vol. 12, n^{o} 2, pp. 4440-4486, https://arxiv.org/abs/1612.04740 , https://hal.archives-ouvertes.fr/hal-01416704

[10] R. LEBLOND, F. PEDREGOSA, S. LACOSTE-JULIEN. *Improved asynchronous parallel optimization analysis for stochastic incremental methods*, in "Journal of Machine Learning Research (JMLR)", 2018, https://hal.inria.fr/hal-01950558

[11] T. LELIEVRE, L. PILLAUD-VIVIEN, J. REYGNER. *Central Limit Theorem for stationary Fleming–Viot particle systems in finite spaces*, in "ALEA : Latin American Journal of Probability and Mathematical Statistics", September 2018, vol. 15, pp. 1163-1182, https://arxiv.org/abs/1806.04490 [*DOI :* 10.30757/ALEA.v15-43], https://hal-enpc.archives-ouvertes.fr/hal-01812120

[12] J. LIN, A. RUDI, L. ROSASCO, V. CEVHER. *Optimal rates for spectral algorithms with least-squares regression over Hilbert spaces*, in "Applied and Computational Harmonic Analysis", October 2018, https://hal.inria.fr/hal-01958890

[13] T. SCHATZ, F. BACH, E. DUPOUX. *Evaluating automatic speech recognition systems as quantitative models of cross-lingual phonetic category perception*, in "Journal of the Acoustical Society of America", May 2018, vol. 143, n^{o} 5, pp. EL372 - EL378 [*DOI :* 10.1121/1.5037615], https://hal.archives-ouvertes.fr/hal-01888735

International Conferences with Proceedings

[14] D. BABICHEV, F. BACH. *Constant Step Size Stochastic Gradient Descent for Probabilistic Modeling*, in "UAI 2018 - Conference on Uncertainty in Artificial Intelligence", Monterey, United States, August 2018, https://arxiv.org/abs/1804.05567 , https://hal.inria.fr/hal-01929810

[15] L. CARRATINO, A. RUDI, L. ROSASCO. *Learning with SGD and Random Features*, in "Advances in Neural Information Processing Systems", Montreal, Canada, December 2018, pp. 10213–10224, https://arxiv.org/abs/1807.06343 - Spotlight, https://hal.archives-ouvertes.fr/hal-01958906

[16] R. LEBLOND, J.-B. ALAYRAC, A. OSOKIN, S. LACOSTE-JULIEN. *SeaRNN: Training RNNs with Global-Local Losses*, in "ICLR 2018 : 6th International Conference on Learning Representations", Vancouver, Canada, April 2018, https://hal.inria.fr/hal-01950555

[17] G. LUISE, A. RUDI, M. PONTIL, C. CILIBERTO. *Differential Properties of Sinkhorn Approximation for Learning with Wasserstein Distance*, in "NIPS 2018 - Advances in Neural Information Processing Systems", Montreal, Canada, December 2018, pp. 5864-5874, https://arxiv.org/abs/1805.11897 - 26 pages, 4 figures, https://hal.inria.fr/hal-01958887

[18] A. RUDI, D. CALANDRIELLO, L. CARRATINO, L. ROSASCO. *On Fast Leverage Score Sampling and Optimal Learning*, in "NeurIPS 2018 - Thirty-second Conference on Neural Information Processing Systems", Montreal, Canada, Advances in Neural Information Processing Systems - NIPS-2018, December 2018, vol. 31, pp. 5677–5687, https://arxiv.org/abs/1810.13258 , https://hal.inria.fr/hal-01958879

[19] A. RUDI, C. CILIBERTO, G. M. MARCONI, L. ROSASCO. *Manifold Structured Prediction*, in "NIPS 2018 - Neural Information Processing Systems Conference", Montreal, Canada, Advances in Neural Information Processing Systems, December 2018, vol. 31, pp. 5615-5626, https://arxiv.org/abs/1806.09908 , https://hal.archives-ouvertes.fr/hal-01958900

[20] T. SHPAKOVA, F. BACH, A. OSOKIN. *Marginal Weighted Maximum Log-likelihood for Efficient Learning of Perturb-and-Map models*, in "UAI 2018 - Conference on Uncertainty in Artificial Intelligence 2018", Monterey, United States, August 2018, https://arxiv.org/abs/1811.08725 , https://hal.inria.fr/hal-01939549

[21] A. B. TAYLOR, B. VAN SCOY, L. LESSARD. *Lyapunov Functions for First-Order Methods: Tight Automated Convergence Guarantees*, in "Proceedings of the 35th International Conference on Machine Learning. PMLR 80:4897-4906", Stockholm, Sweden, July 2018, https://arxiv.org/abs/1803.06073 , https://hal.inria.fr/hal-01902068

Conferences without Proceedings

[22] F. BACH. *Efficient Algorithms for Non-convex Isotonic Regression through Submodular Optimization*, in "Advances in Neural Information Processing Systems", Montreal, Canada, December 2018, https://arxiv.org/abs/1707.09157 , https://hal.archives-ouvertes.fr/hal-01569934

[23] A. BEAUGNON, P. CHIFFLIER, F. BACH. *End-to-End Active Learning for Computer Security Experts*, in "KDD Workshop on Interactive Data Exploration and Analytics (IDEA)", Londres, United Kingdom, August 2018, https://hal.archives-ouvertes.fr/hal-01888983

[24] A. BEAUGNON, P. CHIFFLIER, F. BACH. *End-to-End Active Learning for Computer Security Experts*, in "AAAI Workshop on Artificial Intelligence for Cyber Security (AICS)", New Orleans, United States, February 2018, https://hal.archives-ouvertes.fr/hal-01888976

[25] L. CHIZAT, F. BACH. *On the Global Convergence of Gradient Descent for Over-parameterized Models using Optimal Transport*, in "Advances in Neural Information Processing Systems (NIPS)", Montréal, Canada, December 2018, https://arxiv.org/abs/1805.09545 , https://hal.archives-ouvertes.fr/hal-01798792

[26] A. DÉFOSSEZ, N. ZEGHIDOUR, N. USUNIER, L. BOTTOU, F. BACH. *SING: Symbol-to-Instrument Neural Generator*, in "Conference on Neural Information Processing Systems (NIPS)", Montréal, Canada, December 2018, https://arxiv.org/abs/1810.09785 , https://hal.archives-ouvertes.fr/hal-01899949

[27] R. M. GOWER, N. LE ROUX, F. BACH. *Tracking the gradients using the Hessian: A new look at variance reducing stochastic methods*, in "International Conference on Artificial Intelligence and Statistics (AISTATS)", Canary Islands, Spain, 2018, https://arxiv.org/abs/1710.07462 - 17 pages, 2 figures, 1 table [*DOI :* 10.07462], https://hal.archives-ouvertes.fr/hal-01652152

[28] M. E. HALABI, F. BACH, V. CEVHER. *Combinatorial Penalties: Which structures are preserved by convex relaxations?*, in "AISTATS 2018 - 22nd International Conference on Artificial Intelligence and Statistics", Canary Islands, Spain, April 2018, https://arxiv.org/abs/1710.06273 [*DOI :* 10.06273], https://hal.archives-ouvertes.fr/hal-01652151

[29] T. KERDREUX, F. PEDREGOSA, A. D'ASPREMONT. *Frank-Wolfe with Subsampling Oracle*, in "ICML 2018 - 35th International Conference on Machine Learning", Stockholm, Sweden, July 2018, https://arxiv.org/abs/1803.07348 , https://hal.archives-ouvertes.fr/hal-01927391

[30] A. KUNDU, F. BACH, C. BHATTACHARYYA. *Convex optimization over intersection of simple sets: improved convergence rate guarantees via an exact penalty approach*, in "AISTATS 2018 - 22nd International Conference on Artificial Intelligence and Statistics", Canary Islands, Spain, April 2018, https://arxiv.org/abs/1710.06465 [*DOI :* 10.06465], https://hal.archives-ouvertes.fr/hal-01652149

[31] E. PAUWELS, F. BACH, J.-P. VERT. *Relating Leverage Scores and Density using Regularized Christoffel Functions*, in "Neural Information Processing Systems", Montréal, Canada, December 2018, https://hal.archives-ouvertes.fr/hal-01796591

[32] L. PILLAUD-VIVIEN, A. RUDI, F. BACH. *Exponential convergence of testing error for stochastic gradient methods*, in "Conference on Learning Theory (COLT)", Stockholm, Sweden, July 2018, https://arxiv.org/abs/1712.04755 , https://hal.archives-ouvertes.fr/hal-01662278

[33] L. PILLAUD-VIVIEN, A. RUDI, F. BACH. *Statistical Optimality of Stochastic Gradient Descent on Hard Learning Problems through Multiple Passes*, in "Neural Information Processing Systems (NeurIPS)", Montréal, Canada, December 2018, https://arxiv.org/abs/1805.10074 , https://hal.archives-ouvertes.fr/hal-01799116

[34] S. J. REDDI, M. ZAHEER, S. SRA, B. POCZOS, F. BACH, R. SALAKHUTDINOV, A. J. SMOLA. *A Generic Approach for Escaping Saddle points*, in "AISTATS 2018 - 22nd International Conference on Artificial Intelligence and Statistics", Canary Islands, Spain, April 2018, https://arxiv.org/abs/1709.01434 , https://hal.archives-ouvertes.fr/hal-01652150

[35] K. SCAMAN, F. BACH, S. BUBECK, Y. T. LEE, L. MASSOULIÉ. *Optimal Algorithms for Non-Smooth Distributed Optimization in Networks*, in "Advances In Neural Information Processing systems", Montreal, Canada, December 2018, https://arxiv.org/abs/1806.00291 - 17 pages, https://hal.archives-ouvertes.fr/hal-01957013

[36] D. SCIEUR, E. OYALLON, A. D'ASPREMONT, F. BACH. *Nonlinear Acceleration of CNNs*, in "ICLR Workshop track", Vancouver, Canada, April 2018, https://hal.archives-ouvertes.fr/hal-01805251

[37] N. TRIPURANENI, N. FLAMMARION, F. BACH, M. I. JORDAN. *Averaging Stochastic Gradient Descent on Riemannian Manifolds*, in "Computational Learning Theory (COLT)", Stockholm, Sweden, July 2018, https://arxiv.org/abs/1802.09128 - COLT 2018, https://hal.archives-ouvertes.fr/hal-01957015

Other Publications

[38] M. BARRÉ, A. D'ASPREMONT. M^**-Regularized Dictionary Learning*, October 2018, https://arxiv.org/abs/1810.02748 - working paper or preprint [*DOI :* 10.02748], https://hal.archives-ouvertes.fr/hal-01897496

[39] R. BERTHIER, F. BACH, P. GAILLARD. *Gossip of Statistical Observations using Orthogonal Polynomials*, May 2018, https://arxiv.org/abs/1805.08531 - working paper or preprint, https://hal.archives-ouvertes.fr/hal-01797016

[40] R. BOLLAPRAGADA, D. SCIEUR, A. D'ASPREMONT. *Nonlinear Acceleration of Momentum and Primal-Dual Algorithms*, October 2018, https://arxiv.org/abs/1810.04539 - working paper or preprint [*DOI :* 10.04539], https://hal.archives-ouvertes.fr/hal-01893921

[41] L. CHIZAT, F. BACH. *A Note on Lazy Training in Supervised Differentiable Programming*, December 2018, https://arxiv.org/abs/1812.07956 - working paper or preprint, https://hal.inria.fr/hal-01945578

[42] C. CILIBERTO, F. BACH, A. RUDI. *Localized Structured Prediction*, December 2018, https://arxiv.org/abs/1806.02402 - 53 pages, 7 figures, 1 algorithm, https://hal.inria.fr/hal-01958863

[43] A. DIEULEVEUT, A. DURMUS, F. BACH. *Bridging the Gap between Constant Step Size Stochastic Gradient Descent and Markov Chains*, April 2018, https://arxiv.org/abs/1707.06386 - working paper or preprint, https://hal.archives-ouvertes.fr/hal-01565514

[44] Y. DRORI, A. B. TAYLOR. *Efficient First-order Methods for Convex Minimization: a Constructive Approach*, October 2018, https://arxiv.org/abs/1803.05676 - Code available at https://github.com/AdrienTaylor/GreedyMethods, https://hal.inria.fr/hal-01902048

[45] P. GAILLARD, S. GERCHINOVITZ, M. HUARD, G. STOLTZ. *Uniform regret bounds over R^d for the sequential linear regression problem with the square loss*, February 2018, https://arxiv.org/abs/1805.11386 - working paper or preprint, https://hal.archives-ouvertes.fr/hal-01802004

[46] P. GAILLARD, O. WINTENBERGER. *Efficient online algorithms for fast-rate regret bounds under sparsity*, May 2018, https://arxiv.org/abs/1805.09174 - working paper or preprint, https://hal.archives-ouvertes.fr/hal-01798201

[47] H. HENDRIKX, F. BACH, L. MASSOULIÉ. *Accelerated decentralized optimization with local updates for smooth and strongly convex objectives*, October 2018, working paper or preprint, https://hal.inria.fr/hal-01893568

[48] T. KERDREUX, A. D'ASPREMONT, S. POKUTTA. *Restarting Frank-Wolfe*, October 2018, https://arxiv.org/abs/1810.02429 - working paper or preprint [*DOI :* 10.02429], https://hal.archives-ouvertes.fr/hal-01893922

[49] A. NOWAK-VILA, F. BACH, A. RUDI. *Sharp Analysis of Learning with Discrete Losses*, October 2018, https://arxiv.org/abs/1810.06839 - working paper or preprint, https://hal.archives-ouvertes.fr/hal-01893006

[50] D. M. OSTROVSKII, F. BACH. *Finite-sample Analysis of M-estimators using Self-concordance*, October 2018, https://arxiv.org/abs/1810.06838 - working paper or preprint, https://hal.archives-ouvertes.fr/hal-01895127

[51] A. RECANATI, T. KERDREUX, A. D'ASPREMONT. *Reconstructing Latent Orderings by Spectral Clustering*, July 2018, https://arxiv.org/abs/1807.07122 - working paper or preprint, https://hal.archives-ouvertes.fr/hal-01846269

[52] A. RECANATI, N. SERVANT, J.-P. VERT, A. D'ASPREMONT. *Robust Seriation and Applications to Cancer Genomics*, July 2018, https://arxiv.org/abs/1806.00664 - working paper or preprint, https://hal.archives-ouvertes.fr/hal-01851960

[53] E. K. RYU, A. B. TAYLOR, C. BERGELING, P. GISELSSON. *Operator Splitting Performance Estimation: Tight contraction factors and optimal parameter selection*, December 2018, https://arxiv.org/abs/1812.00146 - working paper or preprint, https://hal.inria.fr/hal-01943622

[54] D. SCIEUR, E. OYALLON, A. D'ASPREMONT, F. BACH. *Nonlinear Acceleration of Deep Neural Networks*, May 2018, working paper or preprint, https://hal.archives-ouvertes.fr/hal-01799269

[55] J. TANG, M. GOLBABAEE, F. BACH, M. E. DAVIES. *Structure-Adaptive Accelerated Coordinate Descent*, October 2018, working paper or preprint, https://hal.archives-ouvertes.fr/hal-01889990

[56] T.-H. VU, A. OSOKIN, I. LAPTEV. *Tube-CNN: Modeling temporal evolution of appearance for object detection in video*, January 2019, https://arxiv.org/abs/1812.02619 - 13 pages, 8 figures, technical report, https://hal.archives-ouvertes.fr/hal-01980339