Activity Report 2019

Project-Team SIERRA

Statistical Machine Learning and Parsimony

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

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:

  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]
  Adrien Taylor [Inria, Starting Research Position, from Oct 2019]

Faculty Member

  Marco Cuturi Cameto [École nationale supérieure de l'électronique et de ses applications, Associate Professor, until Sep 2019]

Post-Doctoral Fellows

  Pierre Yves Masse [CTU Prague, Post-Doctoral Fellow]
  Dmitrii Ostrovskii [Inria, Post-Doctoral Fellow, until Jul 2019]
  Yifan Sun [École Normale Supérieure de Paris, Post-Doctoral Fellow, from Sep 2019]
  Adrien Taylor [Inria, Post-Doctoral Fellow, until Sep 2019]

PhD Students

  Dmitry Babichev [Inria, PhD Student, until Feb 2019]
  Mathieu Barre [École Normale Supérieure de Paris, PhD Student]
  Eloise Berthier [DGA, PhD Student, from Sep 2019]
  Raphael Berthier [Inria, PhD Student]
  Margaux Bregere [EDF, PhD Student]
  Vivien Cabannes [Inria, PhD Student, from Sep 2019]
  Alexandre Defossez [Facebook, PhD Student]
  Radu Alexandru Dragomir [École polytechnique, PhD Student]
  Remi Jezequel [Inria, PhD Student, from Sep 2019]
  Thomas Kerdreux [École polytechnique, PhD Student]
  Ulysse Marteau Ferey [Inria, PhD Student, from Sep 2019]
  Gregoire Mialon [Inria, PhD Student]
  Alex Nowak Vila [La Caixa fellowship, PhD Student]
  Loucas Pillaud Vivien [Ministère de l'Ecologie, de l'Energie, du Développement durable et de la Mer, PhD Student]

Tatiana Shpakova [Inria, PhD Student, until Feb 2019]
Remi Leblond [Inria, until Feb 2019]

Technical staff

Loïc Estève [Inria, Engineer]
Gautier Izacard [CNRS, Engineer, from Nov 2019]

Interns and Apprentices

Vivien Cabannes [Univ Vincennes-Saint Denis, until Aug 2019]
Remi Jezequel [École Normale Supérieure de Paris, from Mar 2019 until Aug 2019]
Ulysse Marteau Ferey [École Normale Supérieure de Paris, from Mar 2019 until Aug 2019]
Quentin Rebjock [Inria, from Sep 2019]

Administrative Assistants

Helene Bessin Rousseau [Inria, Administrative Assistant]
Helene Milome [Inria, Administrative Assistant]

Visiting Scientists

Quentin Berthet [University of Cambridge, from Feb 2019 until Apr 2019]
Eduard Gorbunov [Moscow Institute of Physics and Technology, Oct 2019]
Samy Jelassi [Princeton University, from Jun 2019 until Sep 2019]
Ulysse Marteau Ferey [École Normale Supérieure de Paris, until Mar 2019]
Song Mei [Stanford University, from Sep 2019 until Oct 2019]
Geoffrey Negiar [U.C. Berkeley, from May 2019 until Jul 2019]
Anant Raj [M.P.I. Tubingen, from Oct 2019]
Aadirupa Saha [Indian Institute of Technology, Bangalore, from Nov 2019]

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

- Damien Scieur, Prix de thèse PSL-ADELI
- Francis Bach, Prix Jean-Jacques Moreau

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. Fast Gradient Methods for Symmetric Nonnegative Matrix Factorization.

We describe fast gradient methods for solving the symmetric nonnegative matrix factorization problem (SymNMF). We use recent results on non-Euclidean gradient methods and show that the SymNMF problem is smooth relative to a well-chosen Bregman divergence. This approach provides a simple hyperparameter-free method which comes with theoretical convergence guarantees. We also discuss accelerated variants. Numerical experiments on clustering problems show that our algorithm scales well and reaches both state of the art convergence speed and clustering accuracy for SymNMF methods.

7.2. Naive Feature Selection: Sparsity in Naive Bayes.

Due to its linear complexity, naive Bayes classification remains an attractive supervised learning method, especially in very large-scale settings. We propose a sparse version of naive Bayes, which can be used for feature selection. This leads to a combinatorial maximum-likelihood problem, for which we provide an exact solution in the case of binary data, or a bound in the multinomial case. We prove that our bound becomes tight as the marginal contribution of additional features decreases. Both binary and multinomial sparse models are solvable in time almost linear in problem size, representing a very small extra relative cost compared to the classical naive Bayes. Numerical experiments on text data show that the naive Bayes feature selection method is as statistically effective as state-of-the-art feature selection methods such as recursive feature elimination, l1-penalized logistic regression and LASSO, while being orders of magnitude faster. For a large data set, having more than with 1.6 million training points and about 12 million features, and with a non-optimized CPU implementation, our sparse naive Bayes model can be trained in less than 15 seconds.

7.3. Regularity as Regularization: Smooth and Strongly Convex Brenier Potentials in Optimal Transport.

The problem of estimating Wasserstein distances in high-dimensional spaces suffers from the curse of dimensionality: Indeed, ones needs an exponential (w.r.t. dimension) number of samples for the distance between the two samples to be comparable to that between the two measures. Therefore, regularizing the optimal transport (OT) problem is crucial when using Wasserstein distances in machine learning. One of the greatest achievements of the OT literature in recent years lies in regularity theory: one can prove under suitable hypothesis that the OT map between two measures is Lipschitz, or, equivalently when studying 2-Wasserstein distances, that the Brenier convex potential (whose gradient yields an optimal map) is a smooth function. We propose in this work to go backwards, to adopt instead regularity as a regularization tool. We propose algorithms working on discrete measures that can recover nearly optimal transport maps that have small distortion, or, equivalently, nearly optimal Brenier potential that are strongly convex and smooth. For univariate measures, we show that computing these potentials is equivalent to solving an isotonic regression problem under Lipschitz and strong monotonicity constraints. For multivariate measures the problem boils down to a non-convex QCQP problem. We show that this QCQP can be lifted a semidefinite program. Most importantly, these potentials and their gradient can be evaluated on the measures themselves, but can more generally be evaluated on any new point by solving each time a QP. Building on these two formulations we propose practical algorithms to estimate and evaluate transport maps, and illustrate their performance statistically as well as visually on a color transfer task.

7.4. Ranking and synchronization from pairwise measurements via SVD.

Given a measurement graph $G = (V, E)$ and an unknown signal $r \in R^n$, we investigate algorithms for recovering r from pairwise measurements of the form $r_i - r_j; i, j$ in E. This problem arises in a variety of applications, such as ranking teams in sports data and time synchronization of distributed networks. Framed in the context of ranking, the task is to recover the ranking of n teams (induced by r) given a small subset of

noisy pairwise rank offsets. We propose a simple SVD-based algorithmic pipeline for both the problem of time synchronization and ranking. We provide a detailed theoretical analysis in terms of robustness against both sampling sparsity and noise perturbations with outliers, using results from matrix perturbation and random matrix theory. Our theoretical findings are complemented by a detailed set of numerical experiments on both synthetic and real data, showcasing the competitiveness of our proposed algorithms with other state-of-the-art methods.

7.5. Polyak Steps for Adaptive Fast Gradient Methods.

Accelerated algorithms for minimizing smooth strongly convex functions usually require knowledge of the strong convexity parameter mu. In the case of an unknown mu, current adaptive techniques are based on restart schemes. When the optimal value f^* is known, these strategies recover the accelerated linear convergence bound without additional grid search. In this paper we propose a new approach that has the same bound without any restart, using an online estimation of strong convexity parameter. We show the robustness of the Fast Gradient Method when using a sequence of upper bounds on mu. We also present a good candidate for this estimate sequence and detail consistent empirical results.

7.6. An Accelerated Decentralized Stochastic Proximal Algorithm for Finite Sums.

Modern large-scale finite-sum optimization relies on two key aspects: distribution and stochastic updates. For smooth and strongly convex problems, existing decentralized algorithms are slower than modern accelerated variance-reduced stochastic algorithms when run on a single machine, and are therefore not efficient. Centralized algorithms are fast, but their scaling is limited by global aggregation steps that result in communication bottlenecks. In this work, we propose an efficient Accelerated Decentralized stochastic algorithm for Finite Sums named ADFS, which uses local stochastic proximal updates and randomized pairwise communications between nodes. On n machines, ADFS learns from nm samples in the same time it takes optimal algorithms to learn from m samples on one machine. This scaling holds until a critical network size is reached, which depends on communication delays, on the number of samples m, and on the network topology. We provide a theoretical analysis based on a novel augmented graph approach combined with a precise evaluation of synchronization times and an extension of the accelerated proximal coordinate gradient algorithm to arbitrary sampling. We illustrate the improvement of ADFS over state-of-the-art decentralized approaches with experiments.

7.7. On Lazy Training in Differentiable Programming.

In a series of recent theoretical works, it was shown that strongly overparameterized neural networks trained with gradient-based methods could converge exponentially fast to zero training loss, with their parameters hardly varying. In this work, we show that this "lazy training" phenomenon is not specific to overparameterized neural networks, and is due to a choice of scaling, often implicit, that makes the model behave as its linearization around the initialization, thus yielding a model equivalent to learning with positive-definite kernels. Through a theoretical analysis, we exhibit various situations where this phenomenon arises in non-convex optimization and we provide bounds on the distance between the lazy and linearized optimization paths. Our numerical experiments bring a critical note, as we observe that the performance of commonly used non-linear deep convolutional neural networks in computer vision degrades when trained in the lazy regime. This makes it unlikely that "lazy training" is behind the many successes of neural networks in difficult high dimensional tasks.

7.8. Implicit Regularization of Discrete Gradient Dynamics in Linear Neural Networks.

When optimizing over-parameterized models, such as deep neural networks, a large set of parameters can achieve zero training error. In such cases, the choice of the optimization algorithm and its respective

hyper-parameters introduces biases that will lead to convergence to specific minimizers of the objective. Consequently, this choice can be considered as an implicit regularization for the training of over-parametrized models. In this work, we push this idea further by studying the discrete gradient dynamics of the training of a two-layer linear network with the least-squares loss. Using a time rescaling, we show that, with a vanishing initialization and a small enough step size, this dynamics sequentially learns the solutions of a reduced-rank regression with a gradually increasing rank.

7.9. Efficient Primal-Dual Algorithms for Large-Scale Multiclass Classification.

We develop efficient algorithms to train ℓ_1-regularized linear classifiers with large dimensionality d of the feature space, number of classes k, and sample size n. Our focus is on a special class of losses that includes, in particular, the multiclass hinge and logistic losses. Our approach combines several ideas: (i) passing to the equivalent saddle-point problem with a quasi-bilinear objective; (ii) applying stochastic mirror descent with a proper choice of geometry which guarantees a favorable accuracy bound; (iii) devising non-uniform sampling schemes to approximate the matrix products. In particular, for the multiclass hinge loss we propose a sublinear algorithm with iterations performed in $O(d + n + k)$ arithmetic operations.

7.10. Fast and Faster Convergence of SGD for Over-Parameterized Models (and an Accelerated Perceptron).

Modern machine learning focuses on highly expressive models that are able to fit or interpolate the data completely, resulting in zero training loss. For such models, we show that the stochastic gradients of common loss functions satisfy a strong growth condition. Under this condition, we prove that constant step-size stochastic gradient descent (SGD) with Nesterov acceleration matches the convergence rate of the deterministic accelerated method for both convex and strongly-convex functions. We also show that this condition implies that SGD can find a first-order stationary point as efficiently as full gradient descent in non-convex settings. Under interpolation, we further show that all smooth loss functions with a finite-sum structure satisfy a weaker growth condition. Given this weaker condition, we prove that SGD with a constant step-size attains the deterministic convergence rate in both the strongly-convex and convex settings. Under additional assumptions, the above results enable us to prove an $O(1/k^2)$ mistake bound for k iterations of a stochastic perceptron algorithm using the squared-hinge loss. Finally, we validate our theoretical findings with experiments on synthetic and real datasets.

7.11. Globally Convergent Newton Methods for Ill-conditioned Generalized Self-concordant Losses.

In this project, we study large-scale convex optimization algorithms based on the Newton method applied to regularized generalized self-concordant losses, which include logistic regression and softmax regression. We first prove that our new simple scheme based on a sequence of problems with decreasing regularization parameters is provably globally convergent, that this convergence is linear with a constant factor which scales only logarithmically with the condition number. In the parametric setting, we obtain an algorithm with the same scaling than regular first-order methods but with an improved behavior, in particular in ill-conditioned problems. Second, in the non parametric machine learning setting, we provide an explicit algorithm combining the previous scheme with Nyström projection techniques, and prove that it achieves optimal generalization bounds with a time complexity of order $O(nd_{eff}(\lambda))$, a memory complexity of order $O(d_{eff}(\lambda)^2)$ and no dependence on the condition number, generalizing the results known for least-squares regression. Here n is the number of observations and df λ is the associated degrees of freedom. In particular, this is the first large-scale algorithm to solve logistic and softmax regressions in the non-parametric setting with large condition numbers and theoretical guarantees.

7.12. Efficient online learning with kernels for adversarial large scale problems

We are interested in a framework of online learning with kernels for low-dimensional but large-scale and potentially adversarial datasets. We study the computational and theoretical performance of online variations of kernel Ridge regression. Despite its simplicity, the algorithm we study is the first to achieve the optimal regret for a wide range of kernels with a per-round complexity of order n^α with $\alpha < 2$. The algorithm we consider is based on approximating the kernel with the linear span of basis functions. Our contributions is two-fold: 1) For the Gaussian kernel, we propose to build the basis beforehand (independently of the data) through Taylor expansion. For d-dimensional inputs, we provide a (close to) optimal regret of order $O((logn)^{d+1})$ with per-round time complexity and space complexity $O((logn)^{2d})$. This makes the algorithm a suitable choice as soon as $n \geq e^d$ which is likely to happen in a scenario with small dimensional and large-scale dataset; 2) For general kernels with low effective dimension, the basis functions are updated sequentially in a data-adaptive fashion by sampling Nyström points. In this case, our algorithm improves the computational trade-off known for online kernel regression

7.13. Affine Invariant Covariance Estimation for Heavy-Tailed Distributions

In this work we provide an estimator for the covariance matrix of a heavy-tailed multivariate distribution. We prove that the proposed estimator $\widehat{S}$ admits an *affine-invariant* bound of the form

$$(1 - \epsilon)S \leq \widehat{S}(1 + \epsilon)S$$

in high probability, where S is the unknown covariance matrix, and $\leq$ is the positive semidefinite order on symmetric matrices. The result only requires the existence of fourth-order moments, and allows for $\epsilon = O(\sqrt{k^4 dlog(d/\delta)/n})$ where k^4 is a measure of kurtosis of the distribution, d is the dimensionality of the space, n is the sample size, and $1-\delta$ is the desired confidence level. More generally, we can allow for regularization with level λ, then d gets replaced with the degrees of freedom number. Denoting $cond(S)$ the condition number of S, the computational cost of the novel estimator is $O(d^2 n + d^3 log(cond(S)))$, which is comparable to the cost of the sample covariance estimator in the statistically interesting regime $n \geq d$. We consider applications of our estimator to eigenvalue estimation with relative error, and to ridge regression with heavy-tailed random design.

7.14. Beyond Least-Squares: Fast Rates for Regularized Empirical Risk Minimization through Self-Concordance

We consider learning methods based on the regularization of a convex empirical risk by a squared Hilbertian norm, a setting that includes linear predictors and non-linear predictors through positive-definite kernels. In order to go beyond the generic analysis leading to convergence rates of the excess risk as $O(\sqrt{1/n})$ from n observations, we assume that the individual losses are self-concordant, that is, their third-order derivatives are bounded by their second-order derivatives. This setting includes least-squares, as well as all generalized linear models such as logistic and softmax regression. For this class of losses, we provide a bias-variance decomposition and show that the assumptions commonly made in least-squares regression, such as the source and capacity conditions, can be adapted to obtain fast non-asymptotic rates of convergence by improving the bias terms, the variance terms or both.

7.15. Statistical Estimation of the Poincaré constant and Application to Sampling Multimodal Distributions

Poincaré inequalities are ubiquitous in probability and analysis and have various applications in statistics (concentration of measure, rate of convergence of Markov chains). The Poincaré constant, for which the inequality is tight, is related to the typical convergence rate of diffusions to their equilibrium measure. In this paper, we show both theoretically and experimentally that, given sufficiently many samples of a measure,

we can estimate its Poincaré constant. As a by-product of the estimation of the Poincaré constant, we derive an algorithm that captures a low dimensional representation of the data by finding directions which are difficult to sample. These directions are of crucial importance for sampling or in fields like molecular dynamics, where they are called reaction coordinates. Their knowledge can leverage, with a simple conditioning step, computational bottlenecks by using importance sampling techniques.

7.16. Stochastic first-order methods: non-asymptotic and computer-aided analyses via potential functions

We provide a novel computer-assisted technique for systematically analyzing first-order methods for optimization. In contrast with previous works, the approach is particularly suited for handling sublinear convergence rates and stochastic oracles. The technique relies on semidefinite programming and potential functions. It allows simultaneously obtaining worst-case guarantees on the behavior of those algorithms, and assisting in choosing appropriate parameters for tuning their worst-case performances. The technique also benefits from comfortable tightness guarantees, meaning that unsatisfactory results can be improved only by changing the setting. We use the approach for analyzing deterministic and stochastic first-order methods under different assumptions on the nature of the stochastic noise. Among others, we treat unstructured noise with bounded variance, different noise models arising in over-parametrized expectation minimization problems, and randomized block-coordinate descent schemes.

7.17. Optimal Complexity and Certification of Bregman First-Order Methods

We provide a lower bound showing that the O(1/k) convergence rate of the NoLips method (a.k.a. Bregman Gradient) is optimal for the class of functions satisfying the h-smoothness assumption. This assumption, also known as relative smoothness, appeared in the recent developments around the Bregman Gradient method, where acceleration remained an open issue. On the way, we show how to constructively obtain the corresponding worst-case functions by extending the computer-assisted performance estimation framework of Drori and Teboulle (Mathematical Programming, 2014) to Bregman first-order methods, and to handle the classes of differentiable and strictly convex functions.

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

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 technique builds upon a certain variant of the conjugate gradient method to construct a family of methods such that a) all methods in the family share the same worst-case guarantee as the base conjugate gradient method, and b) the family includes a fixed-step first-order method. We demonstrate the effectiveness of the approach by deriving optimal methods for the smooth and non-smooth cases, including new methods that forego knowledge of the problem parameters at the cost of a one-dimensional line search per iteration, and a universal method for the union of these classes that requires a three-dimensional search per iteration. In the strongly convex case, we show how numerical tools can be used to perform the construction, and show that the resulting method offers an improved worst-case bound compared to Nesterov's celebrated fast gradient method.

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: fondation AXA, "Mécénat scientifique", optimisation & 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

9.2.1. FP7 & H2020 Projects

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, 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.

9.3. International Research Visitors

9.3.1. Visits of International Scientists

- Sebastian Pokutta from TU & Zuse Institute, Berlin, December 2019.
- Critobal Guzman from Universidad Católica de Chile, July 2019.
- Quentin Berthet from University of Cambridge, from Feb 2019 until Apr 2019.
- Eduard Gorbunov from Moscow Institute of Physics and Technology, Oct 2019.
- Song Mei, from Stanford University, from Sep 2019 until Oct 2019.
- Anant Raj, from M.P.I. Tubingen, from Oct 2019.
- Aadirupa Saha, from Indian Institute of Technology, Bangalore, from Nov 2019

10. Dissemination

10.1. Promoting Scientific Activities

- Adrien Taylor, ICCOPT session organizer: *Performance Estimation of First-Order Methods* (with F. Glineur), *Splitting Methods and Applications (Part I)* (with P. Giselsson and E. Ryu), *Splitting Methods and Applications (Part III)* (with P. Giselsson and E. Ryu).

10.1.1. Scientific Events: Organisation

10.1.1.1. Member of the Organizing Committees

- Alexandre d'Aspremont, co-organizer, Les Houches Workshop on Optimization and Machine Learning, March 2019.

10.1.2. Scientific Events: Selection

10.1.2.1. Member of the Conference Program Committees

- Senior Area Chair, NeurIPS conference 2019 (Francis Bach).

10.1.2.2. Reviewer

Most of the team referees for the major machine learning conferences such as NIPS, AISTATS, ICML.

10.1.3. Journal

10.1.3.1. Member of the Editorial Boards

- Alexandre d'Aspremont: SIAM Journal on Optimization, Associate Editor
- Alexandre d'Aspremont: SIAM Journal on the Mathematics of Data Science, Associate Editor
- Alexandre d'Aspremont: Mathematical Programming B, Associate Editor
- Alexandre d'Aspremont: Mathematics of Operations Research, Associate Editor
- Francis Bach: Journal of Machine Learning Research, co-editor-in-chief
- Francis Bach: Information and Inference, Associate Editor
- Francis Bach: Electronic Journal of Statistics, Associate Editor.
- Francis Bach: Mathematical Programming, Associate Editor
- Francis Bach: Foundations of Computational Mathematics, Associate Editor

10.1.3.2. Reviewer - Reviewing Activities

1. Adrien Taylor, reviewer for Automatica.
2. Adrien Taylor, reviewer for SIAM Journal on Numerical Analysis (SINUM).
3. Adrien Taylor, reviewer for SIAM Journal on Optimization (SIOPT).

   4.   Adrien Taylor, reviewer for Mathematical Programming (MPA).

   5.   Adrien Taylor, reviewer for International Conference on Machine Learning 2019 (ICML19).

10.1.4. Invited Talks

- Alexandre d'Aspremont, ICCOPT 2019, Berlin, August 2019.

- Alexandre d'Aspremont, CIMI Workshop, Toulouse, October 2019.

- Alexandre d'Aspremont, France-Germany-Swiss Optimization conference, Nice, September 2019.

- Alexandre d'Aspremont, BIRS Workshop, Oaxaca, Oct. 2019.

- Alexandre d'Aspremont, SAMPTA, Bordeaux, July 2019.

- Francis Bach,Optimization workshop, Les Houches, March 2019.

- Francis Bach, Oberwolfach, Germany, May 2019.

- Francis Bach, AI Global summit, Geneva, Switzerland, May 2019.

- Francis Bach, ETH Data science seminar, Zurich, Switzerland, June 2019.

- Francis Bach, ETH Imaging workshop, Zurich, Switzerland, June 2019.

- Francis BACH, ICIAM invited session, Valencia, Spain, July 2019.

- Francis Bach, Workshop on covariance operators, Germany, Berlin, September 2019.

- Francis Bach, GAMM Workshop COMinDS2019, Berlin, Germany, October 2019.

- Francis Bach, DIMS workshop, Leipzig, Germany, November 2019.

- Francis Bach, Conference on Decision and Control, Nice, December 2019.

- Adrien Taylor, SPOT optimization seminar, Toulouse, February 2019.

- Adrien Taylor, Optimization workshop, Les Houches, March 2019.

- Adrien Taylor, CWI Network & Optimization Seminar, Amsterdam, April 2019.

- Adrien Taylor, Summer school on Optimization (MIPT & HSE), Moscow, June 2019.

- Adrien Taylor, ICCOPT 2019, Berlin, August 2019.

- Alessandro Rudi, *Recent developments on kernel methods*, RKM 2019 (Sept. 2019, London, UK)

- Alessandro Rudi, *Data, Learning and Inference meeting*, DALI 2019 (Sept. 2019, San Sebastian, Spain)

- Alessandro Rudi,*32nd European Meeting of Statisticians*, EMS 2019 (July 2019 , Palermo, Italy)

- Alessandro Rudi,*Applied Inverse Problems Conference*, AIP 2019 (July 2019, Grenoble, France)

- Alessandro Rudi, *Imaging and Machine Learning Conference* (April 2019, Paris, France)

- Alessandro Rudi, *Seminar on the Mathematics of Imaging* (March 2019, Paris, France)

10.1.5. Scientific Expertise

- Alexandre d'Aspremont. Conseil scientifique, Vivienne Investissement.

10.1.6. Research Administration

- Alexandre d'Aspremont. Reponsable scientifique, IRIS PSL, "Sciences des données, données de la science".

- Francis Bach, Deputy Scientific Delegate for Inria Paris research center, member of the Inria Evaluation Committee.

10.2. Teaching - Supervision - Juries

10.2.1. Teaching

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

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

Summer school: Francis Bach, optimization for machine learning, 12 heures, AIMS Master, Kigali, Rwanda

Summer school: Francis Bach, optimization for machine learning, 4 heures, D3S summer school, Ecole Polytechnique.

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

Master: Pierre Gaillard, Alessandro Rudi, Introduction to Machine Learning, 52h, L3, ENS, Paris.

10.2.2. Supervision

PhD in progress : Thomas Kerdreux, New Complexity Bounds for Frank Wolfe, 2017, Alexandre d'Aspremont

PhD in progress : Radu - Dragomir Alexandru, Bregman Gradient Methods, 2018, Alexandre d'Aspremont

PhD in progress : Mathieu Barré, Accelerated Polyak Methods, 2018, Alexandre d'Aspremont

PhD in progress : Grégoire Mialon, Sample Selection Methods, 2018, Alexandre d'Aspremont

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

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

PhD in progress: Alexandre Défossez, supervised by Francis Bach and Léon Bottou (Facebook AI Research).

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

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

PhD in progress: Vivien Cabannes, supervised by Francis Bach and Alessandro Rudi.

PhD in progress: Eloise Berthier, supervised by Francis Bach.

PhD in progress: Theo Ryffel, supervised by Francis Bach and David Pointcheval.

PhD in progress: Margaux Brégère, supervised by Pierre Gaillard and Gilles Stoltz (Université Paris-Sud).

PhD in progress: Rémi Jezequel, supervised by Pierre Gaillard and Alessandro Rudi.

PhD defended: Dmitry Babichev, co-advised by Francis Bach and Anatoly Judistky, defended February 22 2019

PhD defended: Tatiana Shpakova, advised by Francis Bach, defended February 21 2019

10.2.3. Juries

- HdR Pierre Weiss, IMT Toulouse, September 2019 (Alexandre d'Aspremont).
- HDR Rémi Flamary, Université de Nice, November 2019 (Francis Bach).

10.3. Popularization

- Participation to "Fête de la Science" (with the Apprenti Illustrateur).

11. Bibliography

Publications of the year

Doctoral Dissertations and Habilitation Theses

[1] D. BABICHEV. *On Efficient Methods for High-dimensional Statistical Estimation*, PSL Research University, February 2019, https://hal.archives-ouvertes.fr/tel-02433016

[2] T. SHPAKOVA. *On Parameter Learning for Perturb-and-MAP Models*, PSL Research University, February 2019, https://hal.archives-ouvertes.fr/tel-02431640

Articles in International Peer-Reviewed Journals

[3] P. ASKENAZY, F. BACH. *IA et emploi : Une menace artificielle*, in "Pouvoirs - Revue française d'études constitutionnelles et politiques", September 2019, vol. 170, pp. 1-7, https://hal.archives-ouvertes.fr/hal-02343633

[4] M. BRÉGÈRE, P. GAILLARD, Y. GOUDE, G. STOLTZ. *Target Tracking for Contextual Bandits: Application to Demand Side Management*, in "Proceedings of Machine Learning Research", June 2019, vol. 97, pp. 754-763, https://arxiv.org/abs/1901.09532 , https://hal.archives-ouvertes.fr/hal-01994144

[5] Y. DRORI, A. B. TAYLOR. *Efficient First-order Methods for Convex Minimization: a Constructive Approach*, in "Mathematical Programming, Series A", 2019, https://arxiv.org/abs/1803.05676 - Code available at https://github.com/AdrienTaylor/GreedyMethods, forthcoming [*DOI : 10.1007/s10107-019-01410-2*], https://hal.inria.fr/hal-01902048

[6] P. GAILLARD, S. GERCHINOVITZ, M. HUARD, G. STOLTZ. *Uniform regret bounds over R^d for the sequential linear regression problem with the square loss*, in "Proceedings of Machine Learning Research", 2019, vol. 98, pp. 404-432, https://arxiv.org/abs/1805.11386 , https://hal.archives-ouvertes.fr/hal-01802004

[7] J.-L. PEYROT, L. DUVAL, F. PAYAN, L. BOUARD, L. CHIZAT, S. SCHNEIDER, M. ANTONINI. *HexaShrink, an exact scalable framework for hexahedral meshes with attributes and discontinuities: multiresolution rendering and storage of geoscience models*, in "Computational Geosciences", August 2019, vol. 23, n° 4, pp. 723-743, https://arxiv.org/abs/1903.07614 [*DOI : 10.1007/s10596-019-9816-2*], https://hal-ifp.archives-ouvertes.fr/hal-01857997

[8] V. ROULET, N. BOUMAL, A. D'ASPREMONT. *Computational complexity versus statistical performance on sparse recovery problems*, in "Information and Inference", January 2019, https://arxiv.org/abs/1506.03295 [*DOI : 10.1093/IMAIAI/IAY020*], https://hal.archives-ouvertes.fr/hal-02340337

International Conferences with Proceedings

[9] A. BIETTI, G. MIALON, D. CHEN, J. MAIRAL. *A Kernel Perspective for Regularizing Deep Neural Networks*, in "ICML 2019 - 36th International Conference on Machine Learning", Long Beach, United States, Proceedings of Machine Learning Research, June 2019, vol. 97, pp. 664-674, https://arxiv.org/abs/1810.00363 , https://hal.inria.fr/hal-01884632

[10] R. BOLLAPRAGADA, D. SCIEUR, A. D'ASPREMONT. *Nonlinear Acceleration of Momentum and Primal-Dual Algorithms*, in "AISTATS 2019 - 22nd International Conference on Artificial Intelligence and Statistics", Naha, Japan, The 22nd International Conference on Artificial Intelligence and Statistics,, April 2019, vol. 89, https://arxiv.org/abs/1810.04539 [*DOI : 10.04539*], https://hal.archives-ouvertes.fr/hal-01893921

[11] T. KERDREUX, A. D'ASPREMONT, S. POKUTTA. *Restarting Frank-Wolfe*, in "AISTATS 2019 - 22nd International Conference on Artificial Intelligence and Statistics", Naha, Japan, Proceedings of Machine Learning Research, April 2019, vol. 89, https://arxiv.org/abs/1810.02429 [*DOI : 10.02429*], https://hal.archives-ouvertes.fr/hal-01893922

[12] U. MARTEAU-FEREY, F. BACH, A. RUDI. *Globally Convergent Newton Methods for Ill-conditioned Generalized Self-concordant Losses*, in "NeurIPS 2019 - Conference on Neural Information Processing Systems", Vancouver, Canada, December 2019, https://arxiv.org/abs/1907.01771 , https://hal.inria.fr/hal-02169626

[13] T. RYFFEL, E. DUFOUR-SANS, R. GAY, F. BACH, D. POINTCHEVAL. *Partially Encrypted Machine Learning using Functional Encryption*, in "NeurIPS 2019 - Thirty-third Conference on Neural Information Processing Systems", Vancouver, Canada, Advances in Neural Information Processing Systems, December 2019, https://arxiv.org/abs/1905.10214 , https://hal.inria.fr/hal-02357181

[14] A. TAYLOR, F. BACH. *Stochastic first-order methods: non-asymptotic and computer-aided analyses via potential functions*, in "COLT 2019 - Conference on Learning Theory", Phoenix, United States, June 2019, https://arxiv.org/abs/1902.00947 - 12 pages + appendix; code available at https://github.com/AdrienTaylor/Potential-functions-for-first-order-methods, https://hal.inria.fr/hal-02009309

Conferences without Proceedings

[15] P. ABLIN, A. GRAMFORT, J.-F. CARDOSO, F. BACH. *Stochastic algorithms with descent guarantees for ICA*, in "AISTATS 2019 - 22nd International Conference on Artificial Intelligence and Statistics", Naha, Japan, April 2019, https://hal.archives-ouvertes.fr/hal-02372092

[16] L. CHIZAT, E. OYALLON, F. BACH. *On Lazy Training in Differentiable Programming*, in "NeurIPS 2019 - 33rd Conference on Neural Information Processing Systems", Vancouver, Canada, December 2019, https://arxiv.org/abs/1812.07956 , https://hal.inria.fr/hal-01945578

[17] A. GENEVAY, L. CHIZAT, F. BACH, M. CUTURI, G. PEYRÉ. *Sample Complexity of Sinkhorn divergences*, in "AISTATS'19 - 22nd International Conference on Artificial Intelligence and Statistics", Okinawa, Japan, K. CHAUDHURI, M. SUGIYAMA (editors), April 2019, https://arxiv.org/abs/1810.02733 [*DOI :* 10.02733], https://hal.archives-ouvertes.fr/hal-02411822

[18] R. M. GOWER, N. LOIZOU, X. QIAN, A. SAILANBAYEV, E. SHULGIN, P. RICHTÁRIK. *SGD: General Analysis and Improved Rates*, in "International Conference on Machine Learning", Los Angeles, United States, June 2019, https://hal.telecom-paristech.fr/hal-02365318

[19] D. M. OSTROVSKII, A. RUDI. *Affine Invariant Covariance Estimation for Heavy-Tailed Distributions*, in "COLT 2019 - 32nd Annual Conference on Learning Theory", Phoenix, United States, June 2019, https://arxiv.org/abs/1902.03086 , https://hal.archives-ouvertes.fr/hal-02011464

[20] A. PODOSINNIKOVA, A. PERRY, A. WEIN, F. BACH, A. D'ASPREMONT, D. SONTAG. *Overcomplete Independent Component Analysis via SDP*, in "AISTATS 2019 - 22nd International Conference on Artificial Intelligence and Statistics", Naha, Japan, April 2019, https://arxiv.org/abs/1901.08334 - Appears in: Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS 2019). 21 pages, https://hal.archives-ouvertes.fr/hal-02340366

Other Publications

[21] A. ASKARI, A. D'ASPREMONT, L. E. GHAOUI. *Naive Feature Selection: Sparsity in Naive Bayes*, October 2019, https://arxiv.org/abs/1905.09884 - working paper or preprint, https://hal.archives-ouvertes.fr/hal-02340374

[22] D. BABICHEV, D. M. OSTROVSKII, F. BACH. *Efficient Primal-Dual Algorithms for Large-Scale Multiclass Classification*, February 2019, https://arxiv.org/abs/1902.03755 - working paper or preprint, https://hal.archives-ouvertes.fr/hal-02012569

[23] F. BACH. *Max-Plus Matching Pursuit for Deterministic Markov Decision Processes*, June 2019, https://arxiv.org/abs/1906.08524 - working paper or preprint, https://hal.archives-ouvertes.fr/hal-02155865

[24] M. BARRÉ, A. D'ASPREMONT. *Polyak Steps for Adaptive Fast Gradient Method*, October 2019, https://arxiv.org/abs/1906.03056 - working paper or preprint, https://hal.archives-ouvertes.fr/hal-02340373

[25] R. BERTHIER, F. BACH, P. GAILLARD. *Accelerated Gossip in Networks of Given Dimension using Jacobi Polynomial Iterations*, February 2019, https://arxiv.org/abs/1805.08531 - working paper or preprint, https://hal.archives-ouvertes.fr/hal-01797016

[26] A. D'ASPREMONT, M. CUCURINGU, H. TYAGI. *Ranking and synchronization from pairwise measurements via SVD*, October 2019, https://arxiv.org/abs/1906.02746 - 42 pages, 9 figures, https://hal.archives-ouvertes.fr/hal-02340372

[27] R.-A. DRAGOMIR, A. D'ASPREMONT, J. BOLTE. *Quartic First-Order Methods for Low Rank Minimization*, October 2019, https://arxiv.org/abs/1901.10791 - working paper or preprint, https://hal.archives-ouvertes.fr/hal-02340369

[28] R.-A. DRAGOMIR, A. TAYLOR, A. D'ASPREMONT, J. BOLTE. *Optimal Complexity and Certification of Bregman First-Order Methods*, November 2019, https://arxiv.org/abs/1911.08510 - working paper or preprint, https://hal.inria.fr/hal-02384167

[29] A. DÉFOSSEZ, N. USUNIER, L. BOTTOU, F. BACH. *Demucs: Deep Extractor for Music Sources with extra unlabeled data remixed*, September 2019, https://arxiv.org/abs/1909.01174 - working paper or preprint, https://hal.archives-ouvertes.fr/hal-02277338

[30] A. DÉFOSSEZ, N. USUNIER, L. BOTTOU, F. BACH. *Music Source Separation in the Waveform Domain*, November 2019, https://arxiv.org/abs/1911.13254 - working paper or preprint, https://hal.archives-ouvertes.fr/hal-02379796

[31] H. HENDRIKX, F. BACH, L. MASSOULIÉ. *An Accelerated Decentralized Stochastic Proximal Algorithm for Finite Sums*, September 2019, working paper or preprint, https://hal.archives-ouvertes.fr/hal-02280763

[32] R. JÉZÉQUEL, P. GAILLARD, A. RUDI. *Efficient online learning with kernels for adversarial large scale problems*, May 2019, https://arxiv.org/abs/1902.09917 - working paper or preprint, https://hal.archives-ouvertes.fr/hal-02019402

[33] U. MARTEAU-FEREY, D. M. OSTROVSKII, F. BACH, A. RUDI. *Beyond Least-Squares: Fast Rates for Regularized Empirical Risk Minimization through Self-Concordance*, June 2019, https://arxiv.org/abs/1902.03046 - working paper or preprint, https://hal.inria.fr/hal-02011895

[34] G. MIALON, A. D'ASPREMONT, J. MAIRAL. *Screening Data Points in Empirical Risk Minimization via Ellipsoidal Regions and Safe Loss Functions*, December 2019, working paper or preprint, https://hal.archives-ouvertes.fr/hal-02395624

[35] F.-P. PATY, A. D'ASPREMONT, M. CUTURI. *Regularity as Regularization: Smooth and Strongly Convex Brenier Potentials in Optimal Transport*, October 2019, https://arxiv.org/abs/1905.10812 - working paper or preprint, https://hal.archives-ouvertes.fr/hal-02340371

[36] L. PILLAUD-VIVIEN, F. BACH, T. LELIÈVRE, A. RUDI, G. STOLTZ. *Statistical Estimation of the Poincaré constant and Application to Sampling Multimodal Distributions*, November 2019, https://arxiv.org/abs/1910.14564 - working paper or preprint, https://hal.archives-ouvertes.fr/hal-02327453

[37] 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