IN PARTNERSHIP WITH:
CNRS

Ecole normale supérieure de Paris

Activity Report 2017

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:

B9.4.5. - Data science

1. Personnel

Research Scientists

Francis Bach [Team leader, Inria, Senior Researcher, HDR]
Alexandre d'Aspremont [CNRS, Senior Researcher]
Pierre Gaillard [Inria, Researcher]
Alessandro Rudi [Inria, Starting Research Position, from Sep 2017]

Faculty Member

Marco Cuturi Cameto [ENSEA, Associate Professor]

Post-Doctoral Fellows

Lenaic Chizat [Inria, from Dec 2017]
Igor Colin [Institut Telecom ex GET Groupe des Ecoles des Télécommunications]
Pascal Germain [Inria, until Oct 2017]
Robert Gower [Inria, until Aug 2017]
Adrien Taylor [Inria, from Oct 2017]
Federico Vaggi [Ecole Normale Supérieure Paris, until Apr 2017]

PhD Students

Jean-Baptiste Alayrac [Ecole polytechnique, until Sep 2017]
Dmitry Babichev [Inria]
Anaël Bonneton [Ecole Normale Supérieure Paris]
Margaux Bregere [EDF, from Oct 2017]
Alexandre Defossez [Facebook]
Aymeric Dieuleveut [Ecole Normale Supérieure Paris, until Sep 2017]
Christophe Dupuy [Technicolor, until Jun 2017]
Nicolas Flammarion [Ecole Normale Supérieure Lyon, until Aug 2017]
Damien Garreau [Inria, until Aug 2017]
Thomas Kerdreux [Ecole Normale Supérieure Paris, from Oct 2017]
Remi Leblond [Inria]
Loucas Pillaud Vivien [Ministère de l'Ecologie, de l'Energie, du Développement durable et de la Mer, from Sep 2017]
Antoine Recanati [CNRS]

Vincent Roulet [Ecole polytechnique]
Damien Scieur [Inria]
Tatiana Shpakova [Inria]

Technical staff

Fabian Pedregosa [until Apr 2017]

Interns

Brahim Abid [ENSEA, from Apr 2017 until Sep 2017]
Raphael Berthier [Ecole Normale Supérieure Paris, from Sep 2017]
Margaux Bregere [Institut supérieur de l'aéronautique et de l'espace , from Apr 2017 until Oct 2017]
Gauthier Gidel [Ecole Normale Supérieure Paris, until Aug 2017]
Samy Jelassi [Inria, until Jul 2017]
Thomas Kerdreux [Ecole polytechnique, from Apr 2017 until Sep 2017]
Achintya Kundu [Ecole d'ingénieurs, from Sep 2017]
Junqi Tang [from Sep 2017 until Nov 2017]
Cheikh Saliou Toure [Inria, from Apr 2017 until Aug 2017]

Administrative Assistants

Anja Plos [Inria]
Lindsay Polienor [Inria]
Sandrine Verges [Inria]

Visiting Scientists

Marwa El Halabi [Ecole polytechnique, until Apr 2017]
Gauthier Gidel [from Aug 2017]
Lucas Rencker [from Mar 2017 until Sep 2017]
Jonathan Weed [from Mar 2017 until May 2017]
Alfredo Zermini [from Mar 2017 until Jun 2017]

External Collaborators

Christophe Dupuy [from Jul 2017]
Senanayak Karri [until Sep 2017]
Simon Lacoste-Julien
Guillaume Obozinski [Ecole Nationale des Ponts et Chaussées, until Jun 2017]
Balamurugan Palaniappan [École Nationale Supérieure de Techniques Avancées]
Fabian Pedregosa [from Apr 2017 until Aug 2017]
Federico Vaggi [Ecole Normale Supérieure Paris, from May 2017 until Aug 2017]

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. Application Domains

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

- Computer vision: object recognition, object detection, image segmentation, image/video processing, computational photography. In collaboration with the Willow project-team.
- Bioinformatics: cancer diagnosis, protein function prediction, virtual screening. 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. New Software and Platforms

5.1. 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

5.2. 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.

- Contact: Jean-Baptiste Alayrac

6. New Results

6.1. On Structured Prediction Theory with Calibrated Convex Surrogate Losses

In [16], we provide novel theoretical insights on structured prediction in the context of efficient convex surrogate loss minimization with consistency guarantees. For any task loss, we construct a convex surrogate that can be optimized via stochastic gradient descent and we prove tight bounds on the so-called "calibration function" relating the excess surrogate risk to the actual risk. In contrast to prior related work, we carefully monitor the effect of the exponential number of classes in the learning guarantees as well as on the optimization complexity. As an interesting consequence, we formalize the intuition that some task losses make learning harder than others, and that the classical 0-1 loss is ill-suited for general structured prediction.

6.2. Domain-Adversarial Training of Neural Networks

In [18], we introduce a new representation learning approach for domain adaptation, in which data at training and test time come from similar but different distributions. Our approach is directly inspired by the theory on domain adaptation suggesting that, for effective domain transfer to be achieved, predictions must be made based on features that cannot discriminate between the training (source) and test (target) domains. The approach implements this idea in the context of neural network architectures that are trained on labeled data from the source domain and unlabeled data from the target domain (no labeled target-domain data is necessary). As the training progresses, the approach promotes the emergence of features that are (i) discriminative for the main learning task on the source domain and (ii) indiscriminate with respect to the shift between the domains. We show that this adaptation behaviour can be achieved in almost any feed-forward model by augmenting it with few standard layers and a new gradient reversal layer. The resulting augmented architecture can be trained using standard backpropagation and stochastic gradient descent, and can thus be implemented with little effort using any of the deep learning packages. We demonstrate the success of our approach for two distinct classification problems (document sentiment analysis and image classification), where state-of-the-art domain adaptation performance on standard benchmarks is achieved. We also validate the approach for descriptor learning task in the context of person re-identification application.

6.3. Linearly Convergent Randomized Iterative Methods for Computing the Pseudoinverse

In [25], we develop the first stochastic incremental method for calculating the Moore-Penrose pseudoinverse of a real matrix. By leveraging three alternative characterizations of pseudoinverse matrices, we design three methods for calculating the pseudoinverse: two general purpose methods and one specialized to symmetric matrices. The two general purpose methods are proven to converge linearly to the pseudoinverse of any given matrix. For calculating the pseudoinverse of full rank matrices we present two additional specialized methods which enjoy a faster convergence rate than the general purpose methods. We also indicate how to develop randomized methods for calculating approximate range space projections, a much needed tool in inexact Newton type methods or quadratic solvers when linear constraints are present. Finally, we present numerical experiments of our general purpose methods for calculating pseudoinverses and show that our methods greatly outperform the Newton-Schulz method on large dimensional matrices.

6.4. Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance

The Wasserstein distance between two probability measures on a metric space is a measure of closeness with applications in statistics, probability, and machine learning. In [39], we consider the fundamental question of how quickly the empirical measure obtained from n independent samples from μ approaches μ in the Wasserstein distance of any order. We prove sharp asymptotic and finite-sample results for this rate of convergence for general measures on general compact metric spaces. Our finite-sample results show the existence of multi-scale behavior, where measures can exhibit radically different rates of convergence as n grows. Collaboration with Jonathan Weed, Francis Bach)

6.5. Efficient Algorithms for Non-convex Isotonic Regression through Submodular Optimization

In [19], we consider the minimization of submodular functions subject to ordering constraints. We show that this optimization problem can be cast as a convex optimization problem on a space of uni-dimensional measures, with ordering constraints corresponding to first-order stochastic dominance. We propose new discretization schemes that lead to simple and efficient algorithms based on zero-th, first, or higher order oracles; these algorithms also lead to improvements without isotonic constraints. Finally, our experiments show that non-convex loss functions can be much more robust to outliers for isotonic regression, while still leading to an efficient optimization problem.

6.6. Bridging the Gap between Constant Step Size Stochastic Gradient Descent and Markov Chains

In [21], we consider the minimization of an objective function given access to unbiased estimates of its gradient through stochastic gradient descent (SGD) with constant step-size. While the detailed analysis was only performed for quadratic functions, we provide an explicit asymptotic expansion of the moments of the averaged SGD iterates that outlines the dependence on initial conditions, the effect of noise and the step-size, as well as the lack of convergence in the general (non-quadratic) case. For this analysis, we bring tools from Markov chain theory into the analysis of stochastic gradient and create new ones (similar but different from stochastic MCMC methods). We then show that Richardson-Romberg extrapolation may be used to get closer to the global optimum and we show empirical improvements of the new extrapolation scheme.

6.7. AdaBatch: Efficient Gradient Aggregation Rules for Sequential and Parallel Stochastic Gradient Methods

In [22], we study a new aggregation operator for gradients coming from a mini-batch for stochastic gradient (SG) methods that allows a significant speed-up in the case of sparse optimization problems. We call this method AdaBatch and it only requires a few lines of code change compared to regular mini-batch SGD algorithms. We provide a theoretical insight to understand how this new class of algorithms is performing and show that it is equivalent to an implicit per-coordinate rescaling of the gradients, similarly to what Adagrad methods can do. In theory and in practice, this new aggregation allows to keep the same sample efficiency of SG methods while increasing the batch size. Experimentally, we also show that in the case of smooth convex optimization, our procedure can even obtain a better loss when increasing the batch size for a fixed number of samples. We then apply this new algorithm to obtain a parallelizable stochastic gradient method that is synchronous but allows speed-up on par with Hogwild! methods as convergence does not deteriorate with the increase of the batch size. The same approach can be used to make mini-batch provably efficient for variance-reduced SG methods such as SVRG.

6.8. Structure-Adaptive, Variance-Reduced, and Accelerated Stochastic Optimization

In [38], we explore the fundamental structure-adaptiveness of state of the art randomized first order algorithms on regularized empirical risk minimization tasks, where the solution has intrinsic low-dimensional structure (such as sparsity and low-rank). Such structure is often enforced by non-smooth regularization or constraints. We start by establishing the fast linear convergence rate of the SAGA algorithm on non-strongly-convex objectives with convex constraints, via an argument of cone-restricted strong convexity. Then for the composite minimization task with a coordinate-wise separable convex regularization term, we propose and analyse a two stage accelerated coordinate descend algorithm (Two-Stage APCG). We provide the convergence analysis showing that the proposed method has a global convergence in general and enjoys a local accelerated linear convergence rate with respect to the low-dimensional structure of the solution. Then based on this convergence result, we proposed an adaptive variant of the two-stage APCG method which does not need to foreknow the restricted strong convexity beforehand, but estimate it on the fly. In numerical experiments we compare the adaptive two-stage APCG with various state of the art variance-reduced stochastic gradient methods on sparse regression tasks, and demonstrate the effectiveness of our approach.

6.9. Exponential convergence of testing error for stochastic gradient methods

In [31], 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.

6.10. Optimal algorithms for smooth and strongly convex distributed optimization in networks

In [35], we determine the optimal convergence rates for strongly convex and smooth distributed optimization in two settings: centralized and decentralized communications over a network. For centralized (i.e. *master/slave*) algorithms, we show that distributing Nesterov's accelerated gradient descent is optimal and achieves a precision $\varepsilon > 0$ in time $O(\sqrt{\kappa_g}(1+\Delta\tau)\ln(1/\varepsilon))$, where κ_g is the condition number of the (global) function to optimize, Δ is the diameter of the network, and τ (resp. 1) is the time needed to communicate values between two neighbors (resp. perform local computations). For decentralized algorithms based on gossip, we provide the first optimal algorithm, called the *multi-step dual accelerated* (MSDA) method, that achieves a precision $\varepsilon > 0$ in time $O(\sqrt{\kappa_l}(1+\frac{\tau}{\sqrt{\gamma}})\ln(1/\varepsilon))$, where κ_l is the condition number of the local functions and γ is the (normalized) eigengap of the gossip matrix used for communication between nodes. We then verify the efficiency of MSDA against state-of-the-art methods for two problems: least-squares regression and classification by logistic regression.

6.11. Stochastic Composite Least-Squares Regression with convergence rate O(1/n)

In [23], we consider the minimization of composite objective functions composed of the expectation of quadratic functions and an arbitrary convex function. We study the stochastic dual averaging algorithm with a constant step-size, showing that it leads to a convergence rate of $O(1/n)$ without strong convexity assumptions. This thus extends earlier results on least-squares regression with the Euclidean geometry to (a) all convex regularizers and constraints, and (b) all geome-tries represented by a Bregman divergence. This is achieved by a new proof technique that relates stochastic and deterministic recursions.

6.12. Sharpness, Restart and Acceleration

The Łojasiewicz inequality shows that sharpness bounds on the minimum of convex optimization problems hold almost generically. Sharpness directly controls the performance of restart schemes. The constants quantifying error bounds are of course unobservable, but we show in [33] that optimal restart strategies are robust, and searching for the best scheme only increases the complexity by a logarithmic factor compared to the optimal bound. Overall then, restart schemes generically accelerate accelerated methods.

6.13. PAC-Bayes and Domain Adaptation

In [24], we provide two main contributions in PAC-Bayesian theory for domain adaptation where the objective is to learn, from a source distribution, a well-performing majority vote on a different, but related, target distribution. Firstly, we propose an improvement of the previous approach we proposed in Germain et al. (2013), which relies on a novel distribution pseudodistance based on a disagreement averaging, allowing us to derive a new tighter domain adaptation bound for the target risk. While this bound stands in the spirit of common domain adaptation works, we derive a second bound (recently introduced in Germain et al., 2016) that brings a new perspective on domain adaptation by deriving an upper bound on the target risk where the distributions' divergence—expressed as a ratio—controls the trade-off between a source error measure and the target voters' disagreement. We discuss and compare both results, from which we obtain PAC-Bayesian generalization bounds. Furthermore, from the PAC-Bayesian specialization to linear classifiers, we infer two learning algorithms, and we evaluate them on real data.

6.14. Kernel Square-Loss Exemplar Machines for Image Retrieval

Zepeda and Pérez have recently demonstrated the promise of the exemplar SVM (ESVM) as a feature encoder for image retrieval. The paper [6] extends this approach in several directions: We first show that replacing the hinge loss by the square loss in the ESVM cost function significantly reduces encoding time with negligible effect on accuracy. We call this model square-loss exemplar machine, or SLEM. We then introduce a kernelized

SLEM which can be implemented efficiently through low-rank matrix decomposition , and displays improved performance. Both SLEM variants exploit the fact that the negative examples are fixed, so most of the SLEM computational complexity is relegated to an offline process independent of the positive examples. Our experiments establish the performance and computational advantages of our approach using a large array of base features and standard image retrieval datasets.

6.15. Breaking the Nonsmooth Barrier: A Scalable Parallel Method for Composite Optimization

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 [10], 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.

6.16. PAC-Bayesian Analysis for a two-step Hierarchical Multiview Learning Approach

In [15], we study a two-level multiview learning with more than two views under the PAC-Bayesian framework. This approach, sometimes referred as late fusion, consists in learning sequentially multiple view-specific classifiers at the first level, and then combining these view-specific classifiers at the second level. Our main theoretical result is a generalization bound on the risk of the majority vote which exhibits a term of diversity in the predictions of the view-specific classifiers. From this result it comes out that controlling the trade-off between diversity and accuracy is a key element for multiview learning, which complements other results in multiview learning.

6.17. Integration Methods and Accelerated Optimization Algorithms

In [37], we show that accelerated optimization methods can be seen as particular instances of multi-step integration schemes from numerical analysis, applied to the gradient flow equation. In comparison with recent advances in this vein, the differential equation considered here is the basic gradient flow and we show that multi-step schemes allow integration of this differential equation using larger step sizes, thus intuitively explaining acceleration results.

6.18. GANs for Biological Image Synthesis

In [17], we propose a novel application of Generative Adversarial Networks (GAN) to the synthesis of cells imaged by fluorescence microscopy. Compared to natural images, cells tend to have a simpler and more geometric global structure that facilitates image generation. However, the correlation between the spatial pattern of different fluorescent proteins reflects important biological functions, and synthesized images have to capture these relationships to be relevant for biological applications. We adapt GANs to the task at hand and propose new models with casual dependencies between image channels that can generate multi-channel images, which would be impossible to obtain experimentally. We evaluate our approach using two independent techniques and compare it against sensible baselines. Finally, we demonstrate that by interpolating across the latent space we can mimic the known changes in protein localization that occur through time during the cell cycle, allowing us to predict temporal evolution from static images.

6.19. Nonlinear Acceleration of Stochastic Algorithms

Extrapolation methods use the last few iterates of an optimization algorithm to produce a better estimate of the optimum. They were shown to achieve optimal convergence rates in a deterministic setting using simple gradient iterates. In [36], we study extrapolation methods in a stochastic setting, where the iterates are produced by either a simple or an accelerated stochastic gradient algorithm. We first derive convergence bounds for arbitrary, potentially biased perturbations, then produce asymptotic bounds using the ratio between the variance of the noise and the accuracy of the current point. Finally, we apply this acceleration technique to stochastic algorithms such as SGD, SAGA, SVRG and Katyusha in different settings, and show significant performance gains.

6.20. Algorithmic Chaining and the Role of Partial Feedback in Online Nonparametric Learning

In [20], we investigate contextual online learning with nonparametric (Lipschitz) comparison classes under different assumptions on losses and feedback information. For full information feedback and Lipschitz losses, we design the first explicit algorithm achieving the minimax regret rate (up to log factors). In a partial feedback model motivated by second-price auctions, we obtain algorithms for Lipschitz and semi-Lipschitz losses with regret bounds improving on the known bounds for standard bandit feedback. Our analysis combines novel results for contextual second-price auctions with a novel algorithmic approach based on chaining. When the context space is Euclidean, our chaining approach is efficient and delivers an even better regret bound.

6.21. Frank-Wolfe Algorithms for Saddle Point Problems

In [14], we extend the Frank-Wolfe (FW) optimization algorithm to solve constrained smooth convex-concave saddle point (SP) problems. Remarkably, the method only requires access to linear minimization oracles. Leveraging recent advances in FW optimization, we provide the first proof of convergence of a FW-type saddle point solver over polytopes, thereby partially answering a 30 year-old conjecture. We also survey other convergence results and highlight gaps in the theoretical underpinnings of FW-style algorithms. Motivating applications without known efficient alternatives are explored through structured prediction with combinatorial penalties as well as games over matching polytopes involving an exponential number of constraints.

6.22. Convex optimization over intersection of simple sets: improved convergence rate guarantees via an exact penalty approach

In [29], We consider the problem of minimizing a convex function over the intersection of finitely many simple sets which are easy to project onto. This is an important problem arising in various domains such as machine learning. The main difficulty lies in finding the projection of a point in the intersection of many sets. Existing approaches yield an infeasible point with an iteration-complexity of $O(1/\varepsilon^2)$ for nonsmooth problems with no guarantees on the in-feasibility. By reformulating the problem through exact penalty functions, we derive first-order algorithms which not only guarantees that the distance to the intersection is small but also improve the complexity to $O(1/\varepsilon)$ and $O(1/\sqrt{\varepsilon})$ for smooth functions. For composite and smooth problems, this is achieved through a saddle-point reformulation where the proximal operators required by the primal-dual algorithms can be computed in closed form. We illustrate the benefits of our approach on a graph transduction problem and on graph matching. (Collaboration with Achintya Kundu, Francis Bach, Chiranjib Bhattacharyya)

6.23. A Generic Approach for Escaping Saddle points

A central challenge to using first-order methods for optimizing nonconvex problems is the presence of saddle points. First-order methods often get stuck at saddle points, greatly deteriorating their performance. Typically, to escape from saddles one has to use second-order methods. However, most works on second-order methods rely extensively on expensive Hessian-based computations, making them impractical in large-scale settings. To tackle this challenge, we introduce in [32] a generic framework that minimizes Hessian based computations

while at the same time provably converging to second-order critical points. Our framework carefully alternates between a first-order and a second-order subroutine, using the latter only close to saddle points, and yields convergence results competitive to the state-of-the-art. Empirical results suggest that our strategy also enjoys a good practical performance. (Collaboration with Sashank Reddi, Manzil Zaheer, Suvrit Sra, Barnabas Poczos, Ruslan Salakhutdinov, and Alexander Smola)

6.24. Tracking the gradients using the Hessian: A new look at variance reducing stochastic methods

The goal of [26] is to improve variance reducing stochastic methods through better control variates. We first propose a modification of SVRG which uses the Hessian to track gradients over time, rather than to recondition, increasing the correlation of the control variates and leading to faster theoretical convergence close to the optimum. We then propose accurate and computationally efficient approximations to the Hessian, both using a diagonal and a low-rank matrix. Finally, we demonstrate the effectiveness of our method on a wide range of problems.

6.25. Combinatorial Penalties: Which structures are preserved by convex relaxations?

In [28] we consider the homogeneous and the non-homogeneous convex relaxations for combinatorial penalty functions defined on support sets. Our study identifies key differences in the tightness of the resulting relaxations through the notion of the lower combinatorial envelope of a set-function along with new necessary conditions for support identification. We then propose a general adaptive estimator for convex monotone regularizers, and derive new sufficient conditions for support recovery in the asymptotic setting. (Collaboration with Marwa El Halabi, Francis Bach, Volkan Cevher)

6.26. On the Consistency of Ordinal Regression Methods

Many of the ordinal regression models that have been proposed in the literature can be seen as methods that minimize a convex surrogate of the zero-one, absolute, or squared loss functions. A key property that allows to study the statistical implications of such approximations is that of Fisher consistency. Fisher consistency is a desirable property for surrogate loss functions and implies that in the population setting, i.e., if the probability distribution that generates the data were available, then optimization of the surrogate would yield the best possible model. In [3] we will characterize the Fisher consistency of a rich family of surrogate loss functions used in the context of ordinal regression, including support vector ordinal regression, ORBoosting and least absolute deviation. We will see that, for a family of surrogate loss functions that subsumes support vector ordinal regression and ORBoosting, consistency can be fully characterized by the derivative of a real-valued function at zero, as happens for convex margin-based surrogates in binary classification. We also derive excess risk bounds for a surrogate of the absolute error that generalize existing risk bounds for binary classification. Finally, our analysis suggests a novel surrogate of the squared error loss. We compare this novel surrogate with competing approaches on 9 different datasets. Our method shows to be highly competitive in practice, outperforming the least squares loss on 7 out of 9 datasets.

6.27. Iterative hard clustering of features

In [34], we seek to group features in supervised learning problems by constraining the prediction vector coefficients to take only a small number of values. This problem includes non-convex constraints and is solved using projected gradient descent. We prove exact recovery results using restricted eigenvalue conditions. We then extend these results to combine sparsity and grouping constraints, and develop an efficient projection algorithm on the set of grouped and sparse vectors. Numerical experiments illustrate the performance of our algorithms on both synthetic and real data sets.

6.28. Asaga: Asynchronous Parallel Saga

In [9], we describe Asaga, an asynchronous parallel version of the incremental gradient algorithm Saga that enjoys fast linear convergence rates. We highlight 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 proposed "perturbed iterate" framework that resolves it. We thereby prove that Asaga 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 speedup as well as the hardware overhead.

6.29. Sparse Accelerated Exponential Weights

In [8], we consider the stochastic optimization problem where a convex function is minimized observing recursively the gradients. We introduce SAEW, a new procedure that accelerates exponential weights procedures with the slow rate $1/\sqrt{T}$ to procedures achieving the fast rate $1/T$. Under the strong convexity of the risk, we achieve the optimal rate of convergence for approximating sparse parameters in $\mathbb{R}^d$. The acceleration is achieved by using successive averaging steps in an online fashion. The procedure also produces sparse estimators thanks to additional hard threshold steps.

7. Bilateral Contracts and Grants with Industry

7.1. 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/.

7.2. Bilateral Grants with Industry

- A. d'Aspremont: AXA, "mécénat scientifique, chaire Havas-Dauphine", machine learning.
- F. Bach: Gift from Facebook AI Research.

8. Partnerships and Cooperations

8.1. National Initiatives

- A. d'Aspremont: IRIS, PSL "Science des données, données de la science".

8.2. European Initiatives

8.2.1. FP7 & H2020 Projects

- **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 widespread 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.

8.3. International Initiatives

8.3.1. 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: http://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).

8.4. International Research Visitors

8.4.1. Internships

- Marwa El Halabi, from Jan. until Apr. 2017, EPFL, Lausanne, Switzerland
- Jonathan Weed, from Mar. 2017 until May 2017, MIT, US
- Alfredo Zermini, from Mar 2017 until June 2017, University of Surrey, UK
- Billy Tang, visited from Sept. 2017 until Dec. 2017, University of Edimburgh, UK

9. Dissemination

9.1. Promoting Scientific Activities

9.1.1. Scientific Events Organisation

- P. Germain and F. Bach: co-organization of NIPS workshop: "(Almost) 50 Shades of Bayesian Learning: PAC-Bayesian trends and insights" https://bguedj.github.io/nips2017/50shadesbayesian.html
- A. d'Aspremont: co-organization of the workshop: "Optimization and Statistical Learning", Les Houches, France

9.1.1.1. Member of the Organizing Committees

- F. Bach: Senior Area chair for NIPS 2017

9.1.2. Journal

9.1.2.1. Member of the Editorial Boards

- F. Bach: Action Editor, Journal of Machine Learning Research.
- 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.

9.1.3. Invited Talks

- F. Bach: Workshop on Shape, Images and Optimization, Muenster, Germany invited talk, February 2017
- F. Bach: SIAM conference on Optimization, Vancouver, Canada, invited tutorial, May 2017
- F. Bach: LCCC workshop on Large-Scale and Distributed Optimization, Lund, Sweden, invited talk, June 2017
- F. Bach: Summer school on Structured Regularization for High-Dimensional Data Analysis, Paris, invited talk, June 2017
- F. Bach: FOCM Barcelona, two invited talks in special sessions, July 2017
- F. Bach: European Signal Processing conference (EUSIPCO), Kos, Greece, keynote speaker, August 2017
- F. Bach: StatMathAppli 2017, Frejus, mini-course on optimization, September 2017
- F. Bach: 2017 ERNSI Workshop on System Identification, Lyon, invited plenary talk, September 2017
- F. Bach: New-York University, Data science seminar, October 2017
- F. Bach: Workshop on Generative models, parameter learning and sparsity, Cambridge, UK, invited talk, November 2017
- F. Bach: NIPS workshops, two invited talks, Long Beach, CA, December 2017
- A. d'Aspremont: "Regularized Nonlinear Acceleration"
  - GdR MOA, Bordeaux.
  - GdR MEGA, Paris.
  - SIAM OPtimization conference
  - Oxford computational math seminar
  - Alan Turing institute
- A. d'Aspremont: "Sharpness, Restart and Acceleration". Foundations of Computational Mathematics, Barcelona.
- P. Germain: "Generalization of the PAC-Bayesian Theory, and Applications to Semi-Supervised Learning", Modal Seminars, Lille, France, January 2017

- P. Germain: "Theory Driven Domain Adaptation Algorithm", Google Brain TechTalk, Mountain View (CA), USA, April 2017
- P. Gaillard: "Sparse acceleration of exponential weights"
  - Seminar of the SEQUEL project team, Lilles, February 2017
  - 49e Journées Françaises de Statistique, Avignon, Juin 2017
- P. Gaillard: "Obtaining sparse and fast convergence rates online under Bernstein condition", CWI-Inria Workshop, September 2017
- P. Gaillard: "Online nonparametric learning"
  - Cambridge Statistics Seminar, October 2017
  - Statistics Seminar of the University Aix-Marseille, December 2017

9.2. Teaching - Supervision - Juries

9.2.1. Teaching

Master: A. d'Aspremont, "Optimization", 21h, M1, Ecole Normale Supérieure, France

Master: A. d'Aspremont, "Optimization", 21h, M2 (MVA), ENS Cachan, France

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

Master: F. Bach (together with G. Obozinski), "Graphical models", 30h, M2 (MVA), ENS Cachan, France.

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

Master: F. Pedregosa (together with Fajwel Fogel), "Introduction to scikit-learn", M2 (MASH), Université Paris-Dauphine, France.

9.2.2. Supervision

- PhD: Nicolas Flammarion, July 2017, co-directed by Alexandre d'Aspremont and Francis Bach.
- PhD: Aymeric Dieuleveut, September 2017, directed by Francis Bach.
- PhD: Christophe Dupuy, June 2017, directed by Francis Bach.
- PhD: Rafael Rezende, December 2017, Francis Bach, co-advised with Jean Ponce.
- PhD: Vincent Roulet, December 2017, directed by Alexandre d'Aspremont.
- PhD in progress: Damien Scieur, started September 2015, co-directed with Alexandre d'Aspremont and Francis Bach
- PhD in progress: Antoine Recanati, started September 2015, directed by Alexandre d'Aspremont
- PhD in progress: Anaël Bonneton, started December 2014, co-advised by Francis Bach, located in Agence nationale de la sécurité des systèmes d'information (ANSSI).
- PhD in progress: Dmitry Babichev, started September 2015, co-advised by Francis Bach and Anatoly Judistky (Univ. Grenoble).
- PhD in progress: Tatiana Shpakova, started September 2015, advised by Francis Bach.
- PhD in progress: Loucas Pillaud-Vivie, started September 2017, co-directed by Alessandro Rudi and Francis Bach
- PhD in progress: Margaux Brégère, started September 2017, co-advised by Pierre Gaillard, Gilles Stoltz and Yannig Goude (EDF R&D)

9.3. Popularization

- A. d'Aspremont: Paris Science et Data, PSL & Inria.

- A. d'Aspremont: Journée innovation défense
- P. Gaillard: testimony for EDF fellows day

10. Bibliography

Publications of the year

Articles in International Peer-Reviewed Journals

[1] J.-B. ALAYRAC, P. BOJANOWSKI, N. AGRAWAL, J. SIVIC, I. LAPTEV, S. LACOSTE-JULIEN. *Learning from narrated instruction videos*, in "IEEE Transactions on Pattern Analysis and Machine Intelligence", September 2017, vol. XX, https://hal.archives-ouvertes.fr/hal-01580630

[2] F. BACH. *On the Equivalence between Kernel Quadrature Rules and Random Feature Expansions*, in "Journal of Machine Learning Research", 2017, vol. 18, n^{o} 21, pp. 1-38, https://arxiv.org/abs/1502.06800 , https://hal.archives-ouvertes.fr/hal-01118276

[3] F. PEDREGOSA, F. BACH, A. GRAMFORT. *On the Consistency of Ordinal Regression Methods*, in "Journal of Machine Learning Research", 2017, vol. 18, pp. 1 - 35, https://arxiv.org/abs/1408.2327 , https://hal.inria.fr/hal-01054942

[4] N. P. ROUGIER, K. HINSEN, F. ALEXANDRE, T. ARILDSEN, L. BARBA, F. C. Y. BENUREAU, C. T. BROWN, P. DE BUYL, O. CAGLAYAN, A. P. DAVISON, M. A. DELSUC, G. DETORAKIS, A. K. DIEM, D. DRIX, P. ENEL, B. GIRARD, O. GUEST, M. G. HALL, R. N. HENRIQUES, X. HINAUT, K. S. JARON, M. KHAMASSI, A. KLEIN, T. MANNINEN, P. MARCHESI, D. MCGLINN, C. METZNER, O. L. PETCHEY, H. E. PLESSER, T. POISOT, K. RAM, Y. RAM, E. ROESCH, C. ROSSANT, V. ROSTAMI, A. SHIFMAN, J. STACHELEK, M. STIMBERG, F. STOLLMEIER, F. VAGGI, G. VIEJO, J. VITAY, A. VOSTINAR, R. YURCHAK, T. ZITO. *Sustainable computational science: the ReScience initiative*, in "PeerJ Computer Science", 2017, https://arxiv.org/abs/1707.04393 - 8 pages, 1 figure, forthcoming, https://hal.inria.fr/hal-01592078

Articles in Non Peer-Reviewed Journals

[5] A. MEURER, C. SMITH, M. PAPROCKI, O. ČERTÍK, S. KIRPICHEV, M. ROCKLIN, A. KUMAR, S. IVANOV, J. MOORE, S. SINGH, T. RATHNAYAKE, S. VIG, B. GRANGER, R. MULLER, F. BONAZZI, H. GUPTA, S. VATS, F. JOHANSSON, F. PEDREGOSA, M. CURRY, A. TERREL, Š. ROUČKA, A. SABOO, I. FERNANDO, S. KULAL, R. CIMRMAN, A. SCOPATZ. *SymPy: symbolic computing in Python*, in "PeerJ Comput.Sci.", 2017, vol. 3, e103 p. [*DOI :* 10.7717/PEERJ-CS.103], https://hal.archives-ouvertes.fr/hal-01645958

Invited Conferences

[6] R. S. REZENDE, J. ZEPEDA, J. S. PONCE, F. S. BACH, P. PEREZ. *Kernel Square-Loss Exemplar Machines for Image Retrieval*, in "Computer Vision and Pattern Recognition 2017", Honolulu, United States, Computer vision and pattern recognition 2017, July 2017, https://hal.inria.fr/hal-01515224

International Conferences with Proceedings

[7] A. BEAUGNON, P. CHIFFLIER, F. BACH. *ILAB: An Interactive Labelling Strategy for Intrusion Detection*, in "RAID 2017: Research in Attacks, Intrusions and Defenses", Atlanta, United States, September 2017, https://hal.archives-ouvertes.fr/hal-01636299

[8] P. GAILLARD, O. WINTENBERGER. *Sparse Accelerated Exponential Weights*, in "20th International Conference on Artificial Intelligence and Statistics (AISTATS)", Fort Lauderdale, United States, April 2017, https://hal.archives-ouvertes.fr/hal-01376808

[9] R. LEBLOND, F. PEDREGOSA, S. LACOSTE-JULIEN. *Asaga: Asynchronous Parallel Saga*, in "20th International Conference on Artificial Intelligence and Statistics (AISTATS) 2017", Fort Lauderdale, Florida, United States, April 2017, https://hal.inria.fr/hal-01665255

[10] F. PEDREGOSA, R. LEBLOND, S. LACOSTE-JULIEN. *Breaking the Nonsmooth Barrier: A Scalable Parallel Method for Composite Optimization*, in "NIPS 2017 - Thirty-First Annual Conference on Neural Information Processing Systems", Long Beach, United States, December 2017, pp. 1-28, https://arxiv.org/abs/1707.06468 - Appears in Advances in Neural Information Processing Systems 30 (NIPS 2017), 28 pages, https://hal.inria.fr/hal-01638058

[11] F. PEDREGOSA, R. LEBLOND, S. LACOSTE-JULIEN. *Breaking the Nonsmooth Barrier: A Scalable Parallel Method for Composite Optimization*, in "31st Conference on Neural Information Processing Systems (NIPS 2017)", Long Beach, California, United States, December 2017, https://hal.inria.fr/hal-01665260

National Conferences with Proceedings

[12] A. GOYAL, E. MORVANT, P. GERMAIN. *Une borne PAC-Bayésienne en espérance et son extension à l'apprentissage multivues*, in "Conférence Francophone sur l'Apprentissage Automatique (CAp)", Grenoble, France, June 2017, https://hal.archives-ouvertes.fr/hal-01529219

Conferences without Proceedings

[13] A. BEAUGNON, A. HUSSON. *Le Machine Learning confronté aux contraintes opérationnelles des systèmes de détection*, in "SSTIC 2017: Symposium sur la sécurité des technologies de l'information et des communications", Rennes, France, June 2017, pp. 317-346, https://hal.archives-ouvertes.fr/hal-01636303

[14] G. GIDEL, T. JEBARA, S. LACOSTE-JULIEN. *Frank-Wolfe Algorithms for Saddle Point Problems*, in "The 20th International Conference on Artificial Intelligence and Statistics", Fort Lauderdale, Florida, United States, April 2017, https://arxiv.org/abs/1610.07797 , https://hal.archives-ouvertes.fr/hal-01403348

[15] A. GOYAL, E. MORVANT, P. GERMAIN, M.-R. AMINI. *PAC-Bayesian Analysis for a two-step Hierarchical Multiview Learning Approach*, in "European Conference on Machine Learning & Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD)", Skopje, Macedonia, September 2017, Long version : https://arxiv.org/abs/1606.07240, https://hal.archives-ouvertes.fr/hal-01546109

[16] A. OSOKIN, F. BACH, S. LACOSTE-JULIEN. *On Structured Prediction Theory with Calibrated Convex Surrogate Losses*, in "The Thirty-first Annual Conference on Neural Information Processing Systems (NIPS)", Long Beach, United States, December 2017, https://arxiv.org/abs/1703.02403 , https://hal.archives-ouvertes.fr/hal-01611691

[17] A. OSOKIN, A. CHESSEL, R. E. C. SALAS, F. VAGGI. *GANs for Biological Image Synthesis*, in "ICCV 2017 - IEEE International Conference on Computer Vision", Venice, Italy, October 2017, https://arxiv.org/abs/1708.04692 , https://hal.archives-ouvertes.fr/hal-01611692

Scientific Books (or Scientific Book chapters)

[18] Y. GANIN, E. USTINOVA, H. AJAKAN, P. GERMAIN, H. LAROCHELLE, F. LAVIOLETTE, M. MARCHAND, V. LEMPITSKY. *Domain-Adversarial Training of Neural Networks*, in "Domain Adaptation in Computer Vision Applications", G. CSURKA (editor), Advances in Computer Vision and Pattern Recognition, Springer, September 2017 [*DOI :* 10.1007/978-3-319-58347-1], https://hal.archives-ouvertes.fr/hal-01624607

Other Publications

[19] F. BACH. *Efficient Algorithms for Non-convex Isotonic Regression through Submodular Optimization*, July 2017, working paper or preprint, https://hal.archives-ouvertes.fr/hal-01569934

[20] N. CESA-BIANCHI, P. GAILLARD, C. GENTILE, S. GERCHINOVITZ. *Algorithmic Chaining and the Role of Partial Feedback in Online Nonparametric Learning*, June 2017, https://arxiv.org/abs/1702.08211 - This document is the full version of an extended abstract accepted for presentation at COLT 2017., https://hal.archives-ouvertes.fr/hal-01476771

[21] A. DIEULEVEUT, A. DURMUS, F. BACH. *Bridging the Gap between Constant Step Size Stochastic Gradient Descent and Markov Chains*, July 2017, working paper or preprint, https://hal.archives-ouvertes.fr/hal-01565514

[22] A. DÉFOSSEZ, F. BACH. *AdaBatch: Efficient Gradient Aggregation Rules for Sequential and Parallel Stochastic Gradient Methods*, November 2017, working paper or preprint, https://hal.archives-ouvertes.fr/hal-01620513

[23] N. FLAMMARION, F. BACH. *Stochastic Composite Least-Squares Regression with convergence rate O(1/n)*, February 2017, working paper or preprint, https://hal.inria.fr/hal-01472867

[24] P. GERMAIN, A. HABRARD, F. LAVIOLETTE, E. MORVANT. *PAC-Bayes and Domain Adaptation*, July 2017, working paper or preprint, https://hal.archives-ouvertes.fr/hal-01563152

[25] R. M. GOWER, P. RICHTÁRIK. *Linearly Convergent Randomized Iterative Methods for Computing the Pseudoinverse*, January 2017, 28 pages, 10 figures, https://hal.inria.fr/hal-01430489

[26] R. M. GOWER, N. L. ROUX, F. BACH. *Tracking the gradients using the Hessian: A new look at variance reducing stochastic methods*, November 2017, https://arxiv.org/abs/1710.07462 - 17 pages, 2 figures, 1 table [*DOI :* 10.07462], https://hal.archives-ouvertes.fr/hal-01652152

[27] A. GOYAL, E. MORVANT, P. GERMAIN, M.-R. AMINI. *PAC-Bayesian Analysis for a two-step Hierarchical Multiview Learning Approach*, July 2017, working paper or preprint, https://hal.archives-ouvertes.fr/hal-01336260

[28] M. E. HALABI, F. BACH, V. CEVHER. *Combinatorial Penalties: Which structures are preserved by convex relaxations?*, November 2017, working paper or preprint [*DOI :* 10.06273], https://hal.archives-ouvertes.fr/hal-01652151

[29] A. KUNDU, F. BACH, C. BHATTACHARYYA. *Convex optimization over intersection of simple sets: improved convergence rate guarantees via an exact penalty approach*, November 2017, working paper or preprint [*DOI :* 10.06465], https://hal.archives-ouvertes.fr/hal-01652149

[30] R. LEBLOND, J.-B. ALAYRAC, A. OSOKIN, S. LACOSTE-JULIEN. *SEARNN: Training RNNs with global-local losses*, December 2017, https://arxiv.org/abs/1706.04499 - 12 pages, https://hal.inria.fr/hal-01665263

[31] L. PILLAUD-VIVIEN, A. RUDI, F. BACH. *Exponential convergence of testing error for stochastic gradient methods*, December 2017, working paper or preprint, https://hal.archives-ouvertes.fr/hal-01662278

[32] S. J. REDDI, M. ZAHEER, S. SRA, B. POCZOS, F. BACH, R. SALAKHUTDINOV, A. J. SMOLA. *A Generic Approach for Escaping Saddle points*, November 2017, working paper or preprint, https://hal.archives-ouvertes.fr/hal-01652150

[33] V. ROULET, A. D'ASPREMONT. *Sharpness, Restart and Acceleration*, February 2017, working paper or preprint, https://hal.archives-ouvertes.fr/hal-01474362

[34] V. ROULET, F. FOGEL, A. D'ASPREMONT, F. BACH. *Iterative hard clustering of features*, December 2017, working paper or preprint, https://hal.archives-ouvertes.fr/hal-01664964

[35] K. SCAMAN, F. BACH, S. BUBECK, Y. T. LEE, L. MASSOULIÉ. *Optimal algorithms for smooth and strongly convex distributed optimization in networks*, February 2017, working paper or preprint, https://hal.archives-ouvertes.fr/hal-01478317

[36] D. SCIEUR, A. D'ASPREMONT, F. BACH. *Nonlinear Acceleration of Stochastic Algorithms*, October 2017, working paper or preprint, https://hal.archives-ouvertes.fr/hal-01618379

[37] D. SCIEUR, V. ROULET, F. BACH, A. D'ASPREMONT. *Integration Methods and Accelerated Optimization Algorithms*, February 2017, working paper or preprint, https://hal.archives-ouvertes.fr/hal-01474045

[38] J. TANG, F. BACH, M. GOLBABAEE, M. E. DAVIES. *Structure-Adaptive, Variance-Reduced, and Accelerated Stochastic Optimization*, December 2017, working paper or preprint, https://hal.archives-ouvertes.fr/hal-01658487

[39] J. WEED, F. BACH. *Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance*, July 2017, working paper or preprint, https://hal.archives-ouvertes.fr/hal-01555307