IN PARTNERSHIP WITH:
CNRS

Ecole normale supérieure de Paris

Activity Report 2016

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:

1.2.8. - Network security
3.4. - Machine learning and statistics
5.4. - Computer vision
6.2. - Scientific Computing, Numerical Analysis & Optimization
7.1. - Parallel and distributed algorithms
7.3. - Optimization
8.2. - Machine learning

Other Research Topics and Application Domains:

9.4.5. - Data science

1. Members

Research Scientists

Francis Bach [Team leader, Inria, Senior Researcher, HDR]
Alexandre d'Aspremont [CNRS, Senior Researcher, HDR]
Simon Lacoste-Julien [Inria, Starting Research Position, Until Aug 2016]

Engineers

Anton Osokin [Inria]
Fabian Pedregosa [Chaire Havas Dauphine]
Kevin Scaman [Inria, from Nov 2016]

PhD Students

Remi Leblond [Inria]
Jean-Baptiste Alayrac [Ecole Polytechnique]
Dmitry Babichev [Inria]
Anaël Bonneton [ENS Paris]
Alexandre Defossez [CIFRE Facebook]
Aymeric Dieuleveut [ENS Paris]
Christophe Dupuy [CIFRE Technicolor]
Nicolas Flammarion [ENS Lyon]
Damien Garreau [Inria]
Anastasia Podosinnikova [Inria, granted by Microsoft Research]
Antoine Recanati [CNRS]
Vincent Roulet [Ecole Polytechnique]
Damien Scieur [Inria]
Tatiana Shpakova [Inria]

Post-Doctoral Fellows

Amit Bermanis [Inria, until Jul 2016]
Nicolas Boumal [Research in Paris, until Jan 2016]
Pascal Germain [Inria]
Robert Gower [Inria, from Aug 2016]
Balamurugan Palaniappan [Inria]

Federico Vaggi [ENS Paris, from May 2016]

Visiting Scientists

Chiranjib Bhattacharyya [Associated Team Bigfoks2, Nov 2016]
Remi Lajugie [ENS Cachan, until Feb 2016]

Administrative Assistant

Lindsay Polienor

Others

Reza Babanezhad Harikandeh [Inria, from Jun 2016 until Sep 2016]
Fajwel Fogel [Chaire Havas Dauphine, until Mar 2016]
Gauthier Gidel [ENS Paris, from Apr 2016]
Samy Jelassi [Inria, from Feb 2016]
Senanayak Karri [Inria]
Horia Mania [Chaire Havas Dauphine, from May 2016 until Aug 2016]
Guillaume Obozinski [ENPC]

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. DICA : Discrete Independent Component Analysis

FUNCTIONAL DESCRIPTION

Moment Matching for Latent Dirichlet Allocation (LDA) and Discrete Independent Component Analysis (DICA).

The DICA package contains Matlab and C++ (via Matlab mex files) implementations of estimation in the LDA and closely related DICA models.

The implementation consists of two parts. One part contains the efficient implementation for construction of the moment/cumulant tensors, while the other part contains implementations of several so called joint diagonalization type algorithms used for matching the tensors. Any tensor type (see below) can be arbitrarily combined with one of the diagonalization algorithms (see below) leading, in total, to 6 algorithms.

Two types of tensors are considered: (a) the LDA moments and (b) the DICA cumulants. The diagonalization algorithms include: (a) the orthogonal joint diagonalization algorithm based on iterative Jacobi rotations, (b) the spectral algorithm based on two eigen decompositions, and (c) the tensor power method.

- Contact: Anastasia Podosinnikova
- URL: https://github.com/anastasia-podosinnikova/dica

5.2. LinearFW: Implementation of linearly convergent versions of Frank-Wolfe

FUNCTIONAL DESCRIPTION

This is the code to reproduce all the experiments in the NIPS 2015 paper: "On the Global Linear Convergence of Frank-Wolfe Optimization Variants" by Simon Lacoste-Julien and Martin Jaggi, which covers the global linear convergence rate of Frank-Wolfe optimization variants for problems described as in Eq. (1) in the paper. It contains the implementation of Frank-Wolfe, away-steps Frank-Wolfe and pairwise Frank-Wolfe on two applications.

- Contact: Simon Lacoste-Julien
- URL: https://github.com/Simon-Lacoste-Julien/linearFW

5.3. cnn_head_detection: Context-aware CNNs for person head detection

FUNCTIONAL DESCRIPTION

Code for ICCV 2015 paper "Context-aware CNNs for person head detection": Person detection is a key problem for many computer vision tasks. While face detection has reached maturity, detecting people under a full variation of camera view-points, human poses, lighting conditions and occlusions is still a difficult challenge. In this work we focus on detecting human heads in natural scenes. Starting from the recent local R-CNN object detector, we extend it with two types of contextual cues. First, we leverage person-scene relations and propose a Global CNN model trained to predict positions and scales of heads directly from the full image. Second, we explicitly model pairwise relations among objects and train a Pairwise CNN model using a structured-output surrogate loss. The Local, Global and Pairwise models are combined into a joint CNN framework. To train and test our full model, we introduce a large dataset composed of 369,846 human heads annotated in 224,740 movie frames. We evaluate our method and demonstrate improvements of person head detection against several recent baselines in three datasets. We also show improvements of the detection speed provided by our model.

- Contact: Anton Osokin
- URL: https://github.com/aosokin/cnn_head_detection

5.4. Lightning: large-scale linear classification, regression and ranking in Python

FUNCTIONAL DESCRIPTION

Lightning is a Python library for large-scale machine learning. More specifically, the library focuses on linear models for classification, regression and ranking. Lightning is the first project to integrate scikit-learn-contrib, a repository of high-quality projects that follow the same API conventions as scikit-learn. Compared to scikit-learn, the main advantages of lightning are its scalability and its flexibility. Indeed, lightning implements cutting-edge optimization algorithms that allow to train models with millions of samples within seconds on commodity hardware. Furthermore, lightning can leverage prior knowledge thanks to so-called structured penalties, an area of research that has recently found applications in domains as diverse as biology, neuroimaging, finance or text processing. Lightning is available under the 3-clause BSD license at http://contrib.scikit-learn.org/lightning/.

- Contact: Fabian Pedregosa
- URL: http://contrib.scikit-learn.org/lightning/

6. New Results

6.1. Regularized Nonlinear Acceleration

In [34], describe a convergence acceleration technique for generic optimization problems. Our scheme computes estimates of the optimum from a nonlinear average of the iterates produced by any optimization method. The weights in this average are computed via a simple linear system, whose solution can be updated online. This acceleration scheme runs in parallel to the base algorithm, providing improved estimates of the solution on the fly, while the original optimization method is running. Numerical experiments are detailed on classical classification problems.

6.2. Harder, Better, Faster, Stronger Convergence Rates for Least-Squares Regression

In [20], we consider the optimization of a quadratic objective function whose gradients are only accessible through a stochastic oracle that returns the gradient at any given point plus a zero-mean finite variance random error. We present the first algorithm that achieves jointly the optimal prediction error rates for least-squares regression, both in terms of forgetting of initial conditions in $O(1/n^2)$, and in terms of dependence on the noise and dimension d of the problem, as $O(d/n)$. Our new algorithm is based on averaged accelerated regularized gradient descent, and may also be analyzed through finer assumptions on initial conditions and the Hessian matrix, leading to dimension-free quantities that may still be small while the " optimal " terms above are large. In order to characterize the tightness of these new bounds, we consider an application to non-parametric regression and use the known lower bounds on the statistical performance (without computational limits), which happen to match our bounds obtained from a single pass on the data and thus show optimality of our algorithm in a wide variety of particular trade-offs between bias and variance.

6.3. Stochastic Variance Reduction Methods for Saddle-Point Problems

In [12], we consider convex-concave saddle-point problems where the objective functions may be split in many components, and extend recent stochastic variance reduction methods (such as SVRG or SAGA) to provide the first large-scale linearly convergent algorithms for this class of problems which are common in machine learning. While the algorithmic extension is straightforward, it comes with challenges and opportunities: (a) the convex minimization analysis does not apply and we use the notion of monotone operators to prove convergence, showing in particular that the same algorithm applies to a larger class of problems, such as variational inequalities, (b) there are two notions of splits, in terms of functions, or in terms of partial derivatives, (c) the split does need to be done with convex-concave terms, (d) non-uniform sampling is key to an efficient algorithm, both in theory and practice, and (e) these incremental algorithms can be easily accelerated using a simple extension of the “catalyst” framework, leading to an algorithm which is always superior to accelerated batch algorithms.

6.4. Frank-Wolfe Algorithms for Saddle Point Problems

In [26], 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 predic- tion with combinatorial penalties as well as games over matching polytopes involving an exponential number of constraints.

6.5. Minding the Gaps for Block Frank-Wolfe Optimization of Structured SVM

In [10], we propose several improvements on the block-coordinate Frank-Wolfe (BCFW) algorithm from Lacoste-Julien et al. (2013) recently used to optimize the structured support vector machine (SSVM) objective in the context of structured prediction, though it has wider applications. The key intuition behind our improvements is that the estimates of block gaps maintained by BCFW reveal the block suboptimality that can be used as an adaptive criterion. First, we sample objects at each iteration of BCFW in an adaptive non-uniform way via gapbased sampling. Second, we incorporate pairwise and away-step variants of Frank-Wolfe into the block-coordinate setting. Third, we cache oracle calls with a cache-hit criterion based on the block gaps. Fourth, we provide the first method to compute an approximate regularization path for SSVM. Finally, we provide an exhaustive empirical evaluation of all our methods on four structured prediction datasets. The associated SOFTWARE is here: https://github.com/aosokin/gapBCFW

6.6. Asaga: Asynchronous Parallel Saga

In [29], 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.

6.7. Convergence Rate of Frank-Wolfe for Non-Convex Objectives

In [28], we give a simple proof that the Frank-Wolfe algorithm obtains a stationary point at a rate of $O(1/\sqrt{t})$ on non-convex objectives with a Lipschitz continuous gradient. Our analysis is affine invariant and is the first, to the best of our knowledge, giving a similar rate to what was already proven for projected gradient methods (though on slightly different measures of stationarity).

6.8. Highly-Smooth Zero-th Order Online Optimization

The minimization of convex functions which are only available through partial and noisy infor- mation is a key methodological problem in many disciplines. In [3], we consider convex optimization with noisy zero-th order information, that is noisy function evaluations at any desired point. We focus on problems with high degrees of smoothness, such as logistic regression. We show that as opposed to gradient-based algorithms, high-order smoothness may be used to improve estimation rates, with a precise dependence of our upper-bounds on the degree of smoothness. In particular, we show that for infinitely differentiable functions, we recover the same dependence on sample size as gradient-based algorithms, with an extra dimension-dependent factor. This is done for both convex and strongly-convex functions, with finite horizon and anytime algorithms. Finally, we also recover similar results in the online optimization setting.

6.9. Slice Inverse Regression with Score Functions

Non-linear regression and related problems such as non-linear classification are core important tasks in machine learning and statistics. We consider the problem of dimension reduction in non-linear regression, which is often formulated as a non-convex optimization problem.

- We propose score function extensions to sliced inverse regression problems [38], [39], both for the first-order and second-order score functions, which provably improve estimation in the population case over the non-sliced versions; we study finite sample estimators and study their consistency given the exact score functions.
- We propose also to learn the score function as well (using score matching technique [37]) in two steps, i.e., first learning the score function and then learning the effective dimension reduction space, or directly, by solving a convex optimization problem regularized by the nuclear norm.

6.10. Inference and learning for log-supermodular distributions

In [11], we consider log-supermodular models on binary variables, which are probabilistic models with negative log-densities which are submodular. These models provide probabilistic interpretations of common combinatorial optimization tasks such as image segmentation. We make the following contributions:

- We review existing variational bounds for the log-partition function and show that the bound of T. Hazan and T. Jaakkola (On the Partition Function and Random Maximum A-Posteriori Perturbations, Proc. ICML, 2012), based on “perturb-and-MAP” ideas, formally dominates the bounds proposed by J. Djolonga and A. Krause (From MAP to Marginals: Variational Inference in Bayesian Submodular Models, Adv. NIPS, 2014).
- We show that for parameter learning via maximum likelihood the existing bound of J. Djolonga and A. Krause typically leads to a degenerate solution while the one based on “perturb-and-MAP” ideas and logistic samples does not.
- Given that the bound based on “perturb-and-MAP” ideas is an expectation (over our own randomization), we propose to use a stochastic subgradient technique to maximize the lower-bound on the log-likelihood, which can also be extended to conditional maximum likelihood.
- We illustrate our new results on a set of experiments in binary image denoising, where we highlight the flexibility of a probabilistic model for learning with missing data.

6.11. Beyond CCA: Moment Matching for Multi-View Models

In [31], we introduce three novel semi-parametric extensions of probabilistic canonical correlation analysis with identifiability guarantees. We consider moment matching techniques for estimation in these models. For that, by drawing explicit links between the new models and a discrete version of independent component analysis (DICA), we first extend the DICA cumulant tensors to the new discrete version of CCA. By further using a close connection with independent component analysis, we introduce generalized covariance matrices, which can replace the cumulant tensors in the moment matching framework, and, therefore, improve sample complexity and simplify derivations and algorithms significantly. As the tensor power method or orthogonal joint diagonalization are not applicable in the new setting, we use non-orthogonal joint diagonalization techniques for matching the cumulants. We demonstrate performance of the proposed models and estimation techniques on experiments with both synthetic and real datasets.

6.12. PAC-Bayesian Theory Meets Bayesian Inference

In [6], we exhibit a strong link between frequentist PAC-Bayesian bounds and the Bayesian marginal likelihood. That is, for the negative log-likelihood loss function, we show that the minimization of PAC-Bayesian generalization bounds maximizes the Bayesian marginal likelihood. This provides an alternative explanation to the Bayesian Occam’s razor criteria, under the assumption that the data is generated by an *i.i.d.* distribution. Moreover, as the negative log-likelihood is an unbounded loss function, we motivate and propose a PAC-Bayesian theorem tailored for the sub-gamma loss family, and we show that our approach is sound on classical Bayesian linear regression tasks.

6.13. A New PAC-Bayesian Perspective on Domain Adaptation

In [7], we study the issue of PAC-Bayesian domain adaptation: We want to learn, from a source domain, a majority vote model dedicated to a target one. Our theoretical contribution brings a new perspective 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. Our bound suggests that one has to focus on regions where the source data is informative. From this result, we derive a PAC-Bayesian generalization bound, and specialize it to linear classifiers. Then, we infer a learning algorithm and perform experiments on real data.

6.14. PAC-Bayesian Bounds based on the Rényi Divergence

In [13], we propose a simplified proof process for PAC-Bayesian generalization bounds, that allows to divide the proof in four successive inequalities, easing the "customization" of PAC-Bayesian theorems. We also propose a family of PAC-Bayesian bounds based on the Rényi divergence between the prior and posterior distributions, whereas most PAC-Bayesian bounds are based on the Kullback-Leibler divergence. Finally, we present an empirical evaluation of the tightness of each inequality of the simplified proof, for both the classical PAC-Bayesian bounds and those based on the Rényi divergence.

6.15. PAC-Bayesian theorems for multiview learning

In [27], we tackle the issue of multiview learning which aims to take advantages of multiple representations/views of the data. In this context, many machine learning algorithms exist. However, the majority of the theoretical studies focus on learning with exactly two representations. In this paper, we propose a general PAC-Bayesian theory for multiview learning with more than two views. We focus our study to binary classification models that take the form of a majority vote. We derive PAC-Bayesian generalization bounds allowing to consider different relations between empirical and true risks by taking into account a notion of diversity of the voters and views, and that can be naturally extended to semi-supervised learning.

6.16. A spectral algorithm for fast de novo layout of uncorrected long nanopore reads

Seriation is an optimization problem that seeks to reconstruct an ordering between n variables from pairwise similarity information. It can be formulated as a combinatorial problem over permutations and several algorithms have been derived from relaxations of this problem. We make the link between the seriation framework and the task of de novo genome assembly, which consists of reconstructing a whole DNA sequence from small pieces of it that are oversampled so as to cover the full genome. To achieve this task, one has to find the layout of small pieces of DNA sequences (reads). This layout step can be cast as a seriation problem. We show that a spectral algorithm for seriation can be efficiently applied to a genome assembly scheme.

New long read sequencers promise to transform sequencing and genome assembly by producing reads tens of kilobases long. However their high error rate significantly complicates assembly and requires expensive correction steps to layout the reads using standard assembly engines.

We present an original and efficient spectral algorithm to layout the uncorrected nanopore reads, and its seamless integration into a straightforward overlap/layout/consensus (OLC) assembly scheme. The method is shown to assemble Oxford Nanopore reads from several bacterial genomes into good quality ($\sim 99\%$ identity to the reference) genome-sized contigs, while yielding more fragmented assemblies from a *Sacharomyces cerevisiae* reference strain. See software in https://github.com/antrec/spectrassembler.

6.17. Using Deep Learning and Generative Adversarial Networks to Study Large Scale GFP Screens

Fluorescent imaging of GFP tagged proteins is one of the most widely used techniques to view the dynamics of proteins in live cells. By combining it with different perturbations such as RNAi or drug treatments we can understand how cells regulate complex processes such as mitosis or the cell cycle.

However, GFP imaging has certain limitations. There are only a limited number of different fluorescent proteins available, making imaging multiple proteins at the same time very challenging and expensive. Finally, analyzing complex screens can be very challenging: it's not always obvious a-priori what kind of features will predict the phenotypes we are interested in.

We discuss a new approach to studying large scale GFP screens using deep convolutional networks. We show that by using convolutional neural networks, we can greatly outperform traditional feature based approaches at different kind of prediction tasks. The networks learn flexible representations, which are suitable for multiple tasks, such as predicting the localization of Tea1 in fission yeast cells (blue signal, shown in image) in cells where only other proteins are tagged.

We then show that we can use generative adversarial neural networks to learn highly compact latent representations. Those latent representations can then be used to generate new realistic images, allowing us to simulate new phenotypes, and to predict the outcome of new perturbations (joint work between Federico Vaggi, Anton Osokin, Theophile Dalens).

6.18. SymPy: Symbolic computing in Python

SymPy is an open source computer algebra system written in pure Python. It is built with a focus on extensibility and ease of use, through both interactive and programmatic applications. These characteristics have led SymPy to become the standard symbolic library for the scientific Python ecosystem. This paper [30] presents the architecture of SymPy, a description of its features, and a discussion of select domain specific submodules. The supplementary materials provide additional examples and further outline details of the architecture and features of SymPy. As for the software, I am one of the main authors of the lightning machine learning library, that you can include if you want.

6.19. Robust Discriminative Clustering with Sparse Regularizers

Clustering high-dimensional data often requires some form of dimensionality reduction, where clustered variables are separated from "noise-looking" variables. In [24], we cast this problem as finding a low-dimensional projection of the data which is well-clustered. This yields a one-dimensional projection in the simplest situation with two clusters, and extends naturally to a multi-label scenario for more than two clusters. In this paper, (a) we first show that this joint clustering and dimension reduction formulation is equivalent to previously proposed discriminative clustering frameworks, thus leading to convex relaxations of the problem, (b) we propose a novel sparse extension, which is still cast as a convex relaxation and allows estimation in higher dimensions, (c) we propose a natural extension for the multi-label scenario, (d) we provide a new theoretical analysis of the performance of these formulations with a simple probabilistic model, leading to scalings over the form $d = O(\sqrt{n})$ for the affine invariant case and $d = O(n)$ for the sparse case, where n is the number of examples and d the ambient dimension, and finally, (e) we propose an efficient iterative algorithm with running-time complexity proportional to $O(nd^2)$, improving on earlier algorithms which had quadratic complexity in the number of examples.

6.20. Optimal Rates of Statistical Seriation

Given a matrix the seriation problem consists in permuting its rows in such way that all its columns have the same shape, for example, they are monotone increasing. In [23], we propose a statistical approach to this problem where the matrix of interest is observed with noise and study the corresponding minimax rate of estimation of the matrices. Specifically, when the columns are either unimodal or monotone, we show that the least squares estimator is optimal up to logarithmic factors and adapts to matrices with a certain natural structure. Finally, we propose a computationally efficient estimator in the monotonic case and study its performance both theoretically and experimentally. Our work is at the intersection of shape constrained estimation and recent work that involves permutation learning, such as graph denoising and ranking.

6.21. Breaking Sticks and Ambiguities with Adaptive Skip-gram

Recently proposed Skip-gram model is a powerful method for learning high-dimensional word representations that capture rich semantic relationships between words. However, Skip-gram as well as most prior work on learning word representations does not take into account word ambiguity and maintain only single representation per word. Although a number of Skip-gram modifications were proposed to overcome this limitation and learn multi-prototype word representations, they either require a known number of word meanings or learn them using greedy heuristic approaches. In [4], we propose the Adaptive Skip-gram model which is a nonparametric Bayesian extension of Skip-gram capable to automatically learn the required number of representations for all words at desired semantic resolution. We derive efficient online variational learning algorithm for the model and empirically demonstrate its efficiency on word-sense induction task.

6.22. Deep Part-Based Generative Shape Model with Latent Variables

The Shape Boltzmann Machine (SBM) and its multilabel version MSBM [5] have been recently introduced as deep generative models that capture the variations of an object shape. While being more flexible MSBM requires datasets with labeled parts of the objects for training. In [8], we present an algorithm for training MSBM using binary masks of objects and the seeds which approximately correspond to the locations of objects parts. The latter can be obtained from part-based detectors in an unsupervised manner. We derive a latent variable model and an EM-like training procedure for adjusting the weights of MSBM using a deep learning framework. We show that the model trained by our method outperforms SBM in the tasks related to binary shapes and is very close to the original MSBM in terms of quality of multilabel shapes.

6.23. Unsupervised Learning from Narrated Instruction Videos

In [2], we address the problem of automatically learning the main steps to complete a certain task, such as changing a car tire, from a set of narrated instruction videos. The contributions of this paper are three-fold. First, we develop a new unsupervised learning approach that takes advantage of the complementary nature of the input video and the associated narration. The method solves two clustering problems, one in text and one in video, applied one after each other and linked by joint constraints to obtain a single coherent sequence of steps in both modalities. Second, we collect and annotate a new challenging dataset of real-world instruction videos from the Internet. The dataset contains about 800,000 frames for five different tasks that include complex interactions between people and objects, and are captured in a variety of indoor and outdoor settings. Third, we experimentally demonstrate that the proposed method can automatically discover, in an unsupervised manner, the main steps to achieve the task and locate the steps in the input videos. The associated SOFTWARE is here: https://github.com/jalayrac/instructionVideos

6.24. Stochastic Optimization for Large-scale Optimal Transport

Optimal transport (OT) defines a powerful framework to compare probability distributions in a geometrically faithful way. However, the practical impact of OT is still limited because of its computational burden. In [5], we propose a new class of stochastic optimization algorithms to cope with large-scale OT problems. These methods can handle arbitrary distributions (either discrete or continuous) as long as one is able to draw samples from them, which is the typical setup in high-dimensional learning problems. This alleviates the need to discretize these densities, while giving access to provably convergent methods that output the correct distance without discretization error. These algorithms rely on two main ideas: *(a)* the dual OT problem can be re-cast as the maximization of an expectation; *(b)* the entropic regularization of the primal OT problem yields a smooth dual optimization which can be addressed with algorithms that have a provably faster convergence. We instantiate these ideas in three different setups: *(i)* when comparing a discrete distribution to another, we show that incremental stochastic optimization schemes can beat Sinkhorn's algorithm, the current state-of-the-art finite dimensional OT solver; *(ii)* when comparing a discrete distribution to a continuous density, a semi-discrete reformulation of the dual program is amenable to averaged stochastic gradient descent, leading to better performance than approximately solving the problem by discretization ; *(iii)* when dealing with two continuous densities, we propose a stochastic gradient descent over a reproducing kernel Hilbert space (RKHS). This is currently the only known method to solve this problem, apart from computing OT on finite samples. We backup these claims on a set of discrete, semi-discrete and continuous benchmark problems.

6.25. Online but Accurate Inference for Latent Variable Models with Local Gibbs Sampling

We study parameter inference in large-scale latent variable models. We first propose a unified treatment of online inference for latent variable models from a non-canonical exponential family, and draw explicit links between several previously proposed frequentist or Bayesian methods. We then propose a novel inference method for the frequentist estimation of parameters, that adapts MCMC methods to online inference of latent variable models with the proper use of local Gibbs sampling. Then, for latent Dirichlet allocation,we provide an extensive set of experiments and comparisons with existing work, where our new approach outperforms all previously proposed methods. In particular, using Gibbs sampling for latent variable inference is superior to variational inference in terms of test log-likelihoods. Moreover, Bayesian inference through variational methods perform poorly, sometimes leading to worse fits with latent variables of higher dimensionality.

In [22], we focus on methods that make a single pass over the data to estimate parameters. We make the following contributions:

1. We review and compare existing methods for online inference for latent variable models from a non-canonical exponential family, and draw explicit links between several previously proposed frequentist or Bayesian methods. Given the large number of existing methods, our unifying framework allows to understand differences and similarities between all of them.
2. We propose a novel inference method for the frequentist estimation of parameters, that adapts MCMC methods to online inference of latent variable models with the proper use of "local" Gibbs sampling. In our online scheme, we apply Gibbs sampling to the current observation, which is "local", as opposed to "global" batch schemes where Gibbs sampling is applied to the entire dataset.
3. After formulating LDA as a non-canonical exponential family, we provide an extensive set of experiments, where our new approach outperforms all previously proposed methods. In particular, using Gibbs sampling for latent variable inference is superior to variational inference in terms of test log-likelihoods. Moreover, Bayesian inference through variational methods perform poorly, sometimes leading to worse fits with latent variables of higher dimensionality.

6.26. Learning Determinantal Point Processes in Sublinear Time

In [21], we propose a new class of determinantal point processes (DPPs) which can be manipulated for inference and parameter learning in potentially sublinear time in the number of items. This class, based on a specific low-rank factorization of the marginal kernel, is particularly suited to a subclass of continuous DPPs and DPPs defined on exponentially many items. We apply this new class to modelling text documents as sampling a DPP of sentences, and propose a conditional maximum likelihood formulation to model topic proportions, which is made possible with no approximation for our class of DPPs. We present an application to document summarization with a DPP on 2^{500} items.

We make the following contributions:

- We propose a new class of determinantal point processes (DPPs) which is based on a particular low-rank factorization of the marginal kernel. Through the availability of a particular second-moment matrix, the complexity for inference and learning tasks is polynomial in the rank of the factorization and thus often sublinear in the total number of items (with exact likelihood computations).
- As shown in this work, these new DPPs are particularly suited to a subclass of continuous DPPs (infinite number of items), such as on $[0,1]^m$, and DPPs defined on the V-dimensional hypercube, which has 2^V elements.
- We propose a model of documents as sampling a DPP of sentences, and propose a conditional maximum likelihood formulation to model topic proportions. We present an application to document summarization with a DPP on 2^{500} items.

6.27. Decentralized Topic Modelling with Latent Dirichlet Allocation

Privacy preserving networks can be modelled as decentralized networks (e.g., sensors, connected objects, smartphones), where communication between nodes of the network is not controlled by a master or central node. For this type of networks, the main issue is to gather/learn global information on the network (e.g., by optimizing a global cost function) while keeping the (sensitive) information at each node. In this work, we focus on text information that agents do not want to share (e.g., text messages, emails, confidential reports). We use recent advances on decentralized optimization and topic models to infer topics from a graph with limited communication. We propose a method to adapt latent Dirichlet allocation (LDA) model to decentralized optimization and show on synthetic data that we still recover similar parameters and similar performance at each node than with stochastic methods accessing to the whole information in the graph.

In [14], we tackle the non-convex problem of topic modelling, where agents have sensitive text data at their disposal that they can not or do not want to share (e.g., text messages, emails, confidential reports). More precisely, we adapt the particular Latent Dirichlet Allocation (LDA) model to decentralized networks. We combine recent work of [22] on online inference for latent variable models, which adapts online EM with local Gibbs sampling in the case of intractable latent variable models (such as LDA) and recent advances on decentralized optimization.

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.
- A. d’Aspremont: Société Générale - fondation ENS, "mécénat scientifique".
- A. d’Aspremont: Projet EMMA at Institut Louis Bachelier. Collaboration with Euroclear on REPO markets.
- S. Lacoste-Julien (with J. Sivic and I. Laptev in Willow project-team): Google Research Award “Structured Learning from Video and Natural Language”.
- F. Bach: Gift from Facebook AI Research.

8. Partnerships and Cooperations

8.1. European Initiatives

8.1.1. FP7 & H2020 Projects

8.1.1.1. SIPA

Title: Semidefinite Programming with Applications in Statistical Learning

Type: FP7

Instrument: ERC Starting Grant Duration: May 2011 - May 2016 Coordinator: A. d'Aspremont (CNRS)

Abstract: Interior point algorithms and a dramatic growth in computing power have revolutionized optimization in the last two decades. Highly nonlinear problems which were previously thought intractable are now routinely solved at reasonable scales. Semidefinite programs (i.e. linear programs on the cone of positive semidefinite matrices) are a perfect example of this trend: reasonably large, highly nonlinear but convex eigenvalue optimization problems are now solved efficiently by reliable numerical packages. This in turn means that a wide array of new applications for semidefinite programming have been discovered, mimicking the early development of linear programming. To cite only a few examples, semidefinite programs have been used to solve collaborative filtering problems (e.g. make personalized movie recommendations), approximate the solution of combinatorial programs, optimize the mixing rate of Markov chains over networks, infer dependence patterns from multivariate time series or produce optimal kernels in classification problems. These new applications also come with radically different algorithmic requirements. While interior point methods solve relatively small problems with a high precision, most recent applications of semidefinite programming in statistical learning for example form very large-scale problems with comparatively low precision targets, programs for which current algorithms cannot form even a single iteration. This proposal seeks to break this limit on problem size by deriving reliable first-order algorithms for solving large-scale semidefinite programs with a significantly lower cost per iteration, using for example subsampling techniques to considerably reduce the cost of forming gradients. Beyond these algorithmic challenges, the proposed research will focus heavily on applications of convex programming to statistical learning and signal processing theory where optimization and duality results quantify the statistical performance of coding or variable selection algorithms for example. Finally, another central goal of this work will be to produce efficient, customized algorithms for some key problems arising in machine learning and statistics.

8.1.1.2. 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.

8.1.1.3. 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

8.1.1.4. SEQUOIA

Title: Robust algorithms for learning from modern data

Programm: H2020

Type: ERC

Duration: 2017-202

Coordinator: Inria

Inria contact: Francis BACH

8.2. International Initiatives

8.2.1. Inria Associate Teams Not Involved in an Inria International Labs

8.2.1.1. 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).

9. Dissemination

9.1. Promoting Scientific Activities

9.1.1. Scientific Events Organisation

Alexandre d'Aspremont: Workshop preparation for les Houches in Feb. 2016: “Optimization without borders”, to celebrate Y. Nesterov's 60th birthday.

Francis Bach: organization of a workshop at IHES (with S. Arlot and A. Celisse), March 2016.

Francis Bach: co-organization of two NIPS workshops.

9.1.2. Scientific Events Selection

9.1.2.1. Member of the Conference Program Committees

Francis Bach: Area chair for ICML 2016

Simon Lacoste-Julien: Area chair for ICML 2016

Simon Lacoste-Julien: Area chair for NIPS 2016

9.1.3. Journal

9.1.3.1. Member of Editorial Boards

Alexandre d'Aspremont: Associate Editor, SIAM Journal on Optimization (2013-...).

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

F. Bach: Information and Inference, Associate Editor.

F. Bach: SIAM Journal on Imaging Sciences, Associate Editor.

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

9.1.4. Invited Talks

Alexandre d'Aspremont: Regularized Nonlinear Acceleration, BIRS workshop, Oaxaca, October 2016.

Alexandre d'Aspremont: Optimal Affine Invariant Smooth Minimization Algorithms, Institut des hautes études scientifiques, June 2016.

Alexandre d'Aspremont: Optimal Affine Invariant Smooth Minimization Algorithms, Nexus of Information and Computation Theories, Institut Henri Poincaré, March 2016.

Alexandre d'Aspremont: Optimal Affine Invariant Smooth Minimization Algorithms, Workshop on Algorithms and Dynamics for Games and Optimization, Santiago Chile, January 2016.

Francis Bach: Winter School on Signal processing, Bonn, January 2016.

Francis Bach: "Optimization without borders", Les Houches, February 2016.

Francis Bach: Oberwolfach, March 2016.

Francis Bach: Dali meeting, Sestri Levante, Italy, March 2016.

Francis Bach: ETH Computer Science Colloquium, April 2016.

Francis Bach: Workshop San Servolo, May 2016.

Francis Bach: Machine Learning summer school Cadiz, May 2016.

Francis Bach: Summer school, Bangalore, July 2016.

Francis Bach: ICCOPT conference, plenary speaker, August 2016.

Francis Bach: Workshop, Haifa, Septembre 2016.

Francis Bach: Statistics Seminar, Cambridge, October 2016.

Francis Bach: BIRS Oaxaca, October 2016.

Francis Bach: NIPS workshops (three presentations), December 2016.

Damien Garreau: “Consistent multiple change-point detection with kernels”, Group meeting of Geometrica Inria project team, Saclay (February 18, 2016).

Damien Garreau: “Consistent multiple change-point detection with kernels”, Inria Junior Seminar, Paris (March 15, 2016).

Damien Garreau: “Consistent multiple change-point detection with kernels”, Colloque final de l’ANR Calibration, Nice (April 7, 2016).

Damien Garreau: “Consistent multiple change-point detection with kernels”, Colloque Jeunes probabilistes et Statisticiens, Les Houches (April 18, 2016).

Pascal Germain: “A Representation Learning Approach for Domain Adaptation”, Tao Seminars, Université Paris-Sud, Paris, France, March 2016.

Pascal Germain: “A Representation Learning Approach for Domain Adaptation”, Data Intelligence Group Seminars, Université Jean-Monnet, Saint-Étienne, France, March 2016.

Pascal Germain: “Variations on the PAC-Bayesian Bound”, Bayes in Paris Seminar at ENSAE, Paris, France, June 2016.

Pascal Germain: “Variations on the PAC-Bayesian Bound”, Séminaires du département d’informatique et de génie logiciel, Université Laval, Quebec, Canada, July 2016.

Simon Lacoste-Julien: "On the Global Linear Convergence of Frank-Wolfe Optimization Variants", invited talk in the Conic and Polynomial Optimization cluster at ICCOPT 2016, Tokyo, Japan, August 2016..

Simon Lacoste-Julien: "On the Global Linear Convergence of Frank-Wolfe Optimization Variants", invited talk in the Learning and Optimization workshop of DALI meeting, Sestri Levante, Italy, April 2016.

Simon Lacoste-Julien: "Modern Optimization for Structured Machine Learning", CS & OR Department Colloquium, Université de Montréal, Montreal, Canada, February 2016.

Antoine Recanati: Presentation at the group meeting of Mines ParisTech Centre for Computational Biology (CBIO) at Institut Curie, October, 18th 2016.

9.1.5. Leadership within the Scientific Community

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

Alexandre d’Aspremont: Co-scientific director of Master’s program MASH (Mathématiques, Apprentissage et Sciences Humaines), with ENS - Paris Dauphine.

Alexandre d’Aspremont: Scientific committee, programme Gaspard Monge pour l’Optimisation.

9.2. Teaching - Supervision - Juries

9.2.1. Teaching

Master: A. d’Aspremont, M1 course on Optimization: ENS Paris, 21h

Master: A. d’Aspremont, M2 course on Optimization: MVA, ENS Cachan, 21h

Master : F. Bach (together with J.-P. Vert), “Apprentissage statistique”, 35h, M1, Ecole Normale Supérieure.

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

Master : F. Bach , 20h, M2 (Mathématiques de l’aléatoire), Université Paris-Sud.

Mastere (M1): S. Lacoste-Julien, F. Vogel, “Projets informatiques”, 10h, Université de Paris-Dauphine, Master M2: Mathématiques, Apprentissage et Sciences Humaines (MASH)

Master : A. Osokin (together with K. Alahari), “The introduction to discrete optimization”, 30h, M2, Centrale Supélec

Master: Fabian Pedredoga, Machine learning with scikit-learn, Master Mathématiques, Apprentissage et Sciences Humaines (MASH), Paris Dauphine.

9.2.2. Supervision

PhD: Anastasia Podosinnikova, November 2016, co-advised by Francis Bach and Simon Lacoste-Julien

PhD: Thomas Schatz, September 2016, co-advised by and E. Dupoux (ENS, cognitive sciences).

PhD: Sesh Kumar, September 2016, advised by F. Bach.

PhD in progress : Nom du doctorant, titre (provisoire) du mémoire, date du début de la thèse, encadrant(s)

PhD in progress : Jean-Baptiste Alayrac, co-advised by Simon Lacoste-Julien, Josef Sivic and Ivan Laptev, started Sept. 2014.

PhD in progress : Rémi Leblond, advised by Simon Lacoste-Julien, started Sept. 2015.

PhD in progress : Gauthier Gidel, advised by Simon Lacoste-Julien, started Sept. 2016.

PhD in progress : Vincent Roulet, directed by Alexandre d'Aspremont, started as a PhD on Oct. 1 2014.

PhD in progress : Nicolas Flammarion, co-directed by Alexandre d'Aspremont and Francis Bach, started Sept. 2013.

PhD in progress : Damien Scieur, co-directed with Alexandre d'Aspremont and Francis Bach, started Sept. 2015.

PhD in progress : Antoine Recanati, directed by Alexandre d'Aspremont, started Sept. 2015.

PhD in progress: Rafael Rezende, September 2013, F. Bach, co-advised with J. Ponce.

PhD in progress: PhD in progress: Christophe Dupuy, January 2014, co-advised by F. Bach and C. Diot (Technicolor).

PhD in progress: Damien Garreau, September 2014, co-advised by S. Arlot and G. Biau.

PhD in progress: Anaël Bonneton, December 2014, co- advised by F. Bach, located in Agence nationale de la sécurité des systèmes d'information (ANSSI).

PhD in progress: Dmitry Babichev, September 2015, co-advised by F. Bach and A. Judistky (Univ. Grenoble).

PhD in progress: Tatiana Shpakova, September 2015, advised by F. Bach.

9.2.3. Juries

Alexandre d'Aspremont: PhD Committee for Igor Colin, Nov. 2016.

Francis Bach: PhD Committee for Alain Durmus, Dec. 2016.

10. Bibliography

Publications of the year

Articles in International Peer-Reviewed Journals

[1] F. Fogel, I. Waldspurger, A. d'Aspremont. *Phase retrieval for imaging problems*, in "Mathematical Programming Computations", September 2016, vol. 8, n^{o} 3, pp. 311-335, https://hal.archives-ouvertes.fr/hal-00907529

International Conferences with Proceedings

[2] J.-B. Alayrac, P. Bojanowski, N. Agrawal, J. Sivic, I. Laptev, S. Lacoste-Julien. *Unsupervised Learning from Narrated Instruction Videos*, in "CVPR2016 - 29th IEEE Conference on Computer Vision and Pattern Recognition", Las Vegas, United States, June 2016, https://hal.inria.fr/hal-01171193

[3] F. Bach, V. Perchet. *Highly-Smooth Zero-th Order Online Optimization Vianney Perchet*, in "Conference on Learning Theory (COLT)", New York, United States, June 2016, https://hal.archives-ouvertes.fr/hal-01321532

[4] S. Bartunov, D. Kondrashkin, A. Osokin, D. Vetrov. *Breaking Sticks and Ambiguities with Adaptive Skip-gram*, in "Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS)", Cadiz, Spain, May 2016, pp. 130–138, https://hal.archives-ouvertes.fr/hal-01404056

[5] A. Genevay, M. Cuturi, G. Peyré, F. Bach. *Stochastic Optimization for Large-scale Optimal Transport*, in "NIPS 2016 - Thirtieth Annual Conference on Neural Information Processing System", Barcelona, Spain, NIPS (editor), Proc. NIPS 2016, December 2016, https://hal.archives-ouvertes.fr/hal-01321664

[6] P. Germain, F. Bach, A. Lacoste, S. Lacoste-Julien. *PAC-Bayesian Theory Meets Bayesian Inference*, in "Neural Information Processing Systems (NIPS 2016)", Barcelone, Spain, Proceedings of the Neural Information Processing Systems Conference, December 2016, https://hal.archives-ouvertes.fr/hal-01324072

[7] P. Germain, A. Habrard, F. Laviolette, E. Morvant. *A New PAC-Bayesian Perspective on Domain Adaptation*, in "33rd International Conference on Machine Learning (ICML 2016)", New York, NY, United States, Proceedings of the 33rd International Conference on Machine Learning, June 2016, https://hal.archives-ouvertes.fr/hal-01307045

[8] A. Kirillov, M. Gavrikov, E. Lobacheva, A. Osokin, D. Vetrov. *Deep Part-Based Generative Shape Model with Latent Variables*, in "27th British Machine Vision Conference (BMVC 2016)", York, United Kingdom, September 2016, https://hal.archives-ouvertes.fr/hal-01404071

[9] R. Lajugie, P. Bojanowski, P. Cuvillier, S. Arlot, F. Bach. *A weakly-supervised discriminative model for audio-to-score alignment*, in "41st International Conference on Acoustics, Speech, and Signal Processing (ICASSP)", Shanghai, China, Proceedings of the 41st International Conference on Acoustics, Speech, and Signal Processing (ICASSP), March 2016, https://hal.archives-ouvertes.fr/hal-01251018

[10] A. Osokin, J.-B. Alayrac, I. Lukasewitz, P. K. Dokania, S. Lacoste-Julien. *Minding the Gaps for Block Frank-Wolfe Optimization of Structured SVMs*, in "International Conference on Machine Learning (ICML 2016)", New York, United States, 2016, Appears in Proceedings of the 33rd International Conference on Machine Learning (ICML 2016). 31 pages, https://hal.archives-ouvertes.fr/hal-01323727

[11] T. Shpakova, F. Bach. *Parameter Learning for Log-supermodular Distributions*, in "NIPS 2016 - Thirtieth Annual Conference on Neural Information Processing System", Barcelona, Spain, December 2016, https://hal.inria.fr/hal-01354789

Conferences without Proceedings

[12] P. Balamurugan, F. Bach. *Stochastic Variance Reduction Methods for Saddle-Point Problems*, in "Neural Information Processing Systems (NIPS)", Barcelona, Spain, Advances in Neural Information Processing Systems, December 2016, https://hal.archives-ouvertes.fr/hal-01319293

[13] L. Bégin, P. Germain, F. Laviolette, J.-F. Roy. *PAC-Bayesian Bounds based on the Rényi Divergence*, in "International Conference on Artificial Intelligence and Statistics (AISTATS 2016)", Cadiz, Spain, Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, May 2016, https://hal.inria.fr/hal-01384783

[14] I. Colin, C. Dupuy. *Decentralized Topic Modelling with Latent Dirichlet Allocation*, in "NIPS 2016 - 30th Conference on Neural Information Processing Systems", Barcelone, Spain, December 2016, https://hal.archives-ouvertes.fr/hal-01383111

[15] A. GOYAL, E. MORVANT, P. GERMAIN, M.-R. AMINI. *Théorèmes PAC-Bayésiens pour l'apprentissage multi-vues*, in "Conférence Francophone sur l'Apprentissage Automatique (CAp)", Marseille, France, July 2016, https://hal.archives-ouvertes.fr/hal-01329763

[16] L. LANDRIEU, G. OBOZINSKI. *Cut Pursuit: fast algorithms to learn piecewise constant functions*, in "19th International Conference on Artificial Intelligence and Statistics (AISTATS 2016)", Cadix, Spain, May 2016, https://hal.archives-ouvertes.fr/hal-01306786

Research Reports

[17] P. GERMAIN, A. HABRARD, F. LAVIOLETTE, E. MORVANT. *PAC-Bayesian Theorems for Domain Adaptation with Specialization to Linear Classifiers*, Université Jean Monnet, Saint-Étienne (42) ; Département d'Informatique et de Génie Logiciel, Université Laval (Québec) ; ENS Paris ; IST Austria, August 2016, This report is a long version of our paper entitled A PAC-Bayesian Approach for Domain Adaptation with Specialization to Linear Classifiers published in the proceedings of the International Conference on Machine Learning (ICML) 2013. We improved our main results, extended our experiments, and proposed an extension to multisource domain adaptation, https://hal.archives-ouvertes.fr/hal-01134246

Other Publications

[18] D. BABICHEV, F. BACH. *Slice inverse regression with score functions*, October 2016, working paper or preprint, https://hal.inria.fr/hal-01388498

[19] F. BACH. *Submodular Functions: from Discrete to Continous Domains*, February 2016, working paper or preprint, https://hal.archives-ouvertes.fr/hal-01222319

[20] A. DIEULEVEUT, N. FLAMMARION, F. BACH. *Harder, Better, Faster, Stronger Convergence Rates for Least-Squares Regression*, February 2016, working paper or preprint, https://hal.archives-ouvertes.fr/hal-01275431

[21] C. DUPUY, F. BACH. *Learning Determinantal Point Processes in Sublinear Time*, October 2016, Under review for AISTATS 2017, https://hal.archives-ouvertes.fr/hal-01383742

[22] C. DUPUY, F. BACH. *Online but Accurate Inference for Latent Variable Models with Local Gibbs Sampling*, July 2016, Under submission in JMLR, https://hal.inria.fr/hal-01284900

[23] N. FLAMMARION, C. MAO, P. RIGOLLET. *Optimal Rates of Statistical Seriation*, November 2016, V2 corrects an error in Lemma A.1, v3 corrects appendix F on unimodal regression where the bounds now hold with polynomial probability rather than exponential, https://hal.archives-ouvertes.fr/hal-01405738

[24] N. FLAMMARION, B. PALANIAPPAN, F. BACH. *Robust Discriminative Clustering with Sparse Regularizers*, August 2016, working paper or preprint, https://hal.archives-ouvertes.fr/hal-01357666

[25] D. GARREAU, S. ARLOT. *Consistent change-point detection with kernels*, December 2016, working paper or preprint, https://hal.archives-ouvertes.fr/hal-01416704

[26] G. GIDEL, T. JEBARA, S. LACOSTE-JULIEN. *Frank-Wolfe Algorithms for Saddle Point Problems*, October 2016, working paper or preprint, https://hal.archives-ouvertes.fr/hal-01403348

[27] A. GOYAL, E. MORVANT, P. GERMAIN, M.-R. AMINI. *PAC-Bayesian Theorems for Multiview Learning*, November 2016, working paper or preprint, https://hal.archives-ouvertes.fr/hal-01336260

[28] S. LACOSTE-JULIEN. *Convergence Rate of Frank-Wolfe for Non-Convex Objectives*, June 2016, 6 pages, https://hal.inria.fr/hal-01415335

[29] R. LEBLOND, F. PEDREGOSA, S. LACOSTE-JULIEN. *Asaga: Asynchronous Parallel Saga*, December 2016, working paper or preprint, https://hal.archives-ouvertes.fr/hal-01407833

[30] A. MEURER, C. P. SMITH, M. PAPROCKI, O. ČERTÍK, S. B. KIRPICHEV, M. ROCKLIN, A. KUMAR, S. IVANOV, J. K. MOORE, S. SINGH, T. RATHNAYAKE, S. VIG, B. E. GRANGER, R. P. MULLER, F. BONAZZI, H. GUPTA, S. VATS, F. JOHANSSON, F. PEDREGOSA, M. J. CURRY, A. R. TERREL, Š. ROUČKA, A. SABOO, I. FERNANDO, S. KULAL, R. CIMRMAN, A. SCOPATZ. *SymPy: Symbolic computing in Python*, May 2016, working paper or preprint [*DOI :* 10.7287/PEERJ.PREPRINTS.2083V3], https://hal.inria.fr/hal-01404156

[31] A. PODOSINNIKOVA, F. BACH, S. LACOSTE-JULIEN. *Beyond CCA: Moment Matching for Multi-View Models*, March 2016, working paper or preprint, https://hal.inria.fr/hal-01291060

[32] V. ROULET, F. FOGEL, A. D'ASPREMONT, F. BACH. *Learning with Clustering Structure*, October 2016, working paper or preprint, https://hal.archives-ouvertes.fr/hal-01239305

[33] M. SCHMIDT, N. LE ROUX, F. BACH. *Minimizing Finite Sums with the Stochastic Average Gradient*, May 2016, Revision from January 2015 submission. Major changes: updated literature follow and discussion of subsequent work, additional Lemma showing the validity of one of the formulas, somewhat simplified presentation of Lyapunov bound, included code needed for checking proofs rather than the polynomials generated by the code, added error regions to the numerical experiments, https://hal.inria.fr/hal-00860051

[34] D. SCIEUR, A. D'ASPREMONT, F. BACH. *Regularized Nonlinear Acceleration*, November 2016, working paper or preprint, https://hal.archives-ouvertes.fr/hal-01384682

[35] G. SEGUIN, P. BOJANOWSKI, R. LAJUGIE, I. LAPTEV. *Instance-level video segmentation from object tracks*, January 2016, working paper or preprint, https://hal.inria.fr/hal-01255765

[36] K. S. SESH KUMAR, F. BACH. *Active-set Methods for Submodular Minimization Problems*, November 2016, working paper or preprint, https://hal.inria.fr/hal-01161759

References in notes

[37] A. HYVÄRINEN. *Estimation of non-normalized statistical models by score matching*, in "Journal of Machine Learning Research", 2005, vol. 6, pp. 695–709

[38] K.-C. LI. *Sliced Inverse Regression for Dimensional Reduction*, in "Journal of the American Statistical Association", 1991, vol. 86, pp. 316–327

[39] T. M. STOKER. *Consistent estimation of scaled coefficients*, in "Econometrica", 1986, vol. 54, pp. 1461–1481