IN PARTNERSHIP WITH:
CNRS

Ecole normale supérieure de Paris

Activity Report 2013

Project-Team SIERRA

Statistical Machine Learning and Parsimony

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

RESEARCH CENTER
Paris - Rocquencourt

THEME
Optimization, machine learning and statistical methods

Table of contents

Project-Team SIERRA

Keywords: Machine Learning, Statistics, Convex Optimization, Data Mining

Creation of the Project-Team: 2011 January 01.

1. Members

Research Scientists

Francis Bach [Team leader, Inria, HdR]
Sylvain Arlot [CNRS, Researcher]
Simon Lacoste-Julien [Inria, Starting Research position]
Alexandre d'Aspremont [CNRS, Researcher, HdR]

External Collaborators

Toby Hocking [Tokyo Institute of Technology]
Rodolphe Jenatton [CRITEO]
Nicolas Le Roux [CRITEO]
Guillaume Obozinski [ENPC]

PhD Students

Fajwel Fogel [CNRS]
Edouard Grave [Inria, granted by Ecole Polytechnique]
Senanayak Karri [Inria, granted by European Research Council]
Rémi Lajugie [Inria, granted by Fondation d'entreprise EADS]
Anastasia Podosinnikova [Inria, funded by MSR/Inria lab]
Fabian Pedregosa [Inria, joint with PARIETAL team, Saclay]
Florent Couzinié-Devy [ENS cachan, joint with WILLOW team]
Anil Nelakanti [granted by CIFRE]
Thomas Schatz [Univ. Paris VI]
Matthieu Solnon [ENS Paris]
Loic Landrieu [Corps IPEF]

Post-Doctoral Fellows

Mark Schmidt [Inria, until June 30 2013]
Nino Shervashidze [Inria]

Visiting Scientists

Michael Jordan [Inria, from Sep 2012 until Jul 2013]
Bamdev Mishra [Université de Liège]

Administrative Assistant

Lindsay Polienor [Inria]

Others

Matteo Tanzi [Intern from Mar 2013 until May 2013]
Aymeric Dieuleveut [ENS Paris, intern from Apr 2013]
Nicolas Flammarion [ENS Lyon, intern from Apr 2013]
Damien Garreau [ENS Paris, intern from 21 Oct 2013]

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.

2.2. Highlights of the Year

- Visit of Prof. Michael Jordan (U.C. Berkeley) and of his research group.
- Recruitment of two researchers: Alexandre d'Aspremont (DR2 CNRS) and Simon Lacoste-Julien (Inria Starting researcher position).
- Start of a collaboration with Microsoft Research (within the joint MSR/Inria lab).

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 (this is the main topic of the ERC starting investigator grant awarded to F. Bach).

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: objet 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. In collaboration with Telecom Paristech.
- Neuro-imaging: brain-computer interface (fMRI, EEG, MEG). In collaboration with the Parietal project-team.

4.2. Natural Language Processing

This year, our research has focused on new application domains within natural language processing (NLP), with our first two publications in leading conferences in NLP. We have worked on large-scale semantic role labelling (E. Grave, F. Bach, G. Obozinski), where we use syntactic dependency trees and learned representations from large corpora (e.g., 14.7 millions sentences, 310 millions tokens). We also extended our original work on structured sparsity to language models (F. Bach, A. Nelakanti, in collaboration with Xerox), in order to predict a word given *all* previous words, with a potentially infinite feature space organized with structured regularization.

5. Software and Platforms

5.1. SPAMS (SPArse Modeling Software)

Participants: Jean-Paul Chieze [correspondent], Guillaume Obozinski [correspondent].

SPAMS (SPArse Modeling Software) is an optimization toolbox for solving various sparse estimation problems: dictionary learning and matrix factorization, solving sparse decomposition prob- lems, solving structured sparse decomposition problems. It is developped by Julien Mairal (former Willow PhD student, co-advised by F. Bach and J. Ponce), with the collaboration of Francis Bach (Inria), Jean Ponce (Ecole Normale Supérieure), Guillermo Sapiro (University of Minnesota), Rodolphe Jenatton (Inria) and Guillaume Obozinski (Inria). It is coded in C++ with a Matlab interface. This year, interfaces for R and Python have been developed by Jean-Paul Chieze (engineer Inria). Currently 650 downloads and between 1500 and 2000 page visits per month. See http://spams-devel.gforge.inria.fr/.

5.2. BCFWstruct

Participants: Simon Lacoste-Julien [correspondent], Mark Schmidt.

BCFWstruct is a Matlab implementation of the Block-Coordinate Frank-Wolfe solver for Structural SVMs. See the ICML 2013 paper with the same name.

Participants outside of Sierra: Martin Jaggi (Centre de Mathématiques Appliquées, Ecole Polytechnique); Patrick Pletscher (Machine Learning Laboratory, ETH Zurich)

5.3. SAG

Participant: Mark Schmidt [correspondent].

SAG: Minimizing Finite Sums with the Stochastic Average Gradient.

The SAG code contains C implements (via Matlab mex files) of the stochastic average gradient (SAG) method detailed below, as well as several related methods, for the problem of L2-regularized logistic regression with a finite training set.

The specific methods available in the package are: SGD: The stochastic gradient method with (user-supplied) step-sizes, (optional) projection step, and (optional) (weighted-)averaging. ASGD: A variant of the above code that supports less features, but efficiently implements uniform averaging on sparse data sets. PCD: A basic primal coordinate descent method with step sizes set according the (user-supplied) Lipschitz constants. DCA: A dual coordinate ascent method with a numerical high-accuracy line-search. SAG: The stochastic average gradient method with a (user-supplied) constant step size. SAGlineSearch: The stochastic average gradient method with the line-search described in the paper. SAG-LipschitzLS: The stochastic average gradient method with the line-search and adaptive non-uniform sampling strategy described in the paper.

5.4. fMRI

Participant: Fabian Pedregosa [correspondent].

We showed that HRF estimation improves sensitivity of fMRI encoding and decoding models and propose a new approach for the estimation of Hemodynamic Response Functions from fMRI data. This is an implementation of the methods described in the paper.

6. New Results

6.1. Block-Coordinate Frank-Wolfe Optimization for Structural SVMs

Participants: Simon Lacoste-Julien [correspondent], Mark Schmidt.

Collaboration with: Martin Jaggi (Centre de Mathématiques Appliquées, Ecole Polytechnique), Patrick Pletscher (Machine Learning Laboratory, ETH Zurich).

In [16] we propose a randomized block-coordinate variant of the classic Frank-Wolfe algorithm for convex optimization with block-separable constraints. Despite its lower iteration cost, we show that it achieves a similar convergence rate in duality gap as the full Frank-Wolfe algorithm. We also show that, when applied to the dual structural support vector machine (SVM) objective, it yields an online algorithm that has the same low iteration complexity as primal stochastic subgradient methods. However, unlike stochastic subgradient methods, the block-coordinate Frank-Wolfe algorithm allows us to compute the optimal step-size and yields a computable duality gap guarantee. Our experiments indicate that this simple algorithm outperforms competing structural SVM solvers.

6.2. Minimizing Finite Sums with the Stochastic Average Gradient.

Participants: Mark Schmidt [correspondent], Nicolas Le Roux, Francis Bach.

In [32] we propose the stochastic average gradient (SAG) method for optimizing the sum of a finite number of smooth convex functions. Like stochastic gradient (SG) methods, the SAG method's iteration cost is independent of the number of terms in the sum. However, by incorporating a memory of previous gradient values the SAG method achieves a faster convergence rate than black-box SG methods. The convergence rate is improved from $O(1/\sqrt{k})$ to $O(1/k)$ in general, and when the sum is strongly-convex the convergence rate is improved from the sub-linear $O(1/k)$ to a linear convergence rate of the form $O(\rho^k)$ for $\rho < 1$. Further, in many cases the convergence rate of the new method is also faster than black-box deterministic gradient methods, in terms of the number of gradient evaluations. Numerical experiments indicate that the new algorithm often dramatically outperforms existing SG and deterministic gradient methods.

The primary contribution of this work is the analysis of a new algorithm that we call the *stochastic average gradient* (SAG) method, a randomized variant of the incremental aggregated gradient (IAG) method of [43]. The SAG method has the low iteration cost of SG methods, but achieves the convergence rates stated above for the FG method. The SAG iterations take the form

$$x^{k+1} = x^k - \frac{\alpha_k}{n} \sum_{i=1}^{n} y_i^k, \tag{1}$$

where at each iteration a random index i_k is selected and we set $y_i^k = f_i'(x^k)$ if $i = i_k$, and y_i^{k-1} otherwise. That is, like the FG method, the step incorporates a gradient with respect to each function. But, like the SG method, each iteration only computes the gradient with respect to a single example and the cost of the iterations is independent of n. Despite the low cost of the SAG iterations, we show in this paper that with a constant step-size *the SAG iterations have an $O(1/k)$ convergence rate for convex objectives and a linear convergence rate for strongly-convex objectives*, like the FG method. That is, by having access to i_k and by keeping a *memory* of the most recent gradient value computed for each index i, this iteration achieves a faster convergence rate than is possible for standard SG methods. Further, in terms of effective passes through the data, we will also see that for many problems the convergence rate of the SAG method is also faster than is possible for standard FG methods.

6.3. Fast Convergence of Stochastic Gradient Descent under a Strong Growth Condition

Participants: Mark Schmidt [correspondent], Nicolas Le Roux [correspondent].

In [33] we consider optimizing a function smooth convex function f that is the average of a set of differentiable functions f_i, under the assumption considered by [87] and [90] that the norm of each gradient f_i' is bounded by a linear function of the norm of the average gradient f'. We show that under these assumptions the basic stochastic gradient method with a sufficiently-small constant step-size has an $O(1/k)$ convergence rate, and has a linear convergence rate if g is strongly-convex.

We write our problem

$$\min_{x \in \mathbb{R}^P} f(x) := \frac{1}{N} \sum_{i=1}^{N} f_i(x), \tag{2}$$

where we assume that f is convex and its gradient f' is Lipschitz-continuous with constant L, meaning that for all x and y we have

$$||f'(x) - f'(y)|| \leq L||x - y||.$$

If f is twice-differentiable, these assumptions are equivalent to assuming that the eigenvalues of the Hessian $f''(x)$ are bounded between 0 and L for all x.

Deterministic gradient methods for problems of this form use the iteration

$$x_{k+1} = x_k - \alpha_k f'(x_k), \tag{3}$$

for a sequence of step sizes α_k. In contrast, *stochastic gradient* methods use the iteration

$$x_{k+1} = x_k - \alpha_k f_i'(x_k), \tag{4}$$

for an individual data sample i selected uniformly at random from the set $\{1, 2, \cdots, N\}$.

The stochastic gradient method is appealing because the cost of its iterations is *independent of* N. However, in order to guarantee convergence stochastic gradient methods require a decreasing sequence of step sizes $\{\alpha_k\}$ and this leads to a slower convergence rate. In particular, for convex objective functions the stochastic gradient method with a decreasing sequence of step sizes has an expected error on iteration k of $O(1/\sqrt{k})$ [78], meaning that

$$\mathbb{E}[f(x_k)] - f(x^*) = O(1/\sqrt{k}).$$

In contrast, the deterministic gradient method with a *constant* step size has a smaller error of $O(1/k)$ [79]. The situation is more dramatic when f is *strongly* convex, meaning that

$$f(y) \geq f(x) + \langle f'(x), y - x\rangle + \frac{\mu}{2}||y - x||^2, \tag{5}$$

for all x and y and some $\mu > 0$. For twice-differentiable functions, this is equivalent to assuming that the eigenvalues of the Hessian are bounded below by μ. For strongly convex objective functions, the stochastic gradient method with a decreasing sequence of step sizes has an error of $O(1/k)$ [77] while the deterministic method with a constant step size has an *linear* convergence rate. In particular, the deterministic method satisfies

$$f(x_k) - f(x^*) \leq \rho^k [f(x_0) - f(x^*)],$$

for some $\rho < 1$ [71].

We show that if the individual gradients $f_i'(x_k)$ satisfy a certain strong growth condition relative to the full gradient $f'(x_k)$, the stochastic gradient method with a sufficiently small constant step size achieves (in expectation) the convergence rates stated above for the deterministic gradient method.

6.4. Non-strongly-convex smooth stochastic approximation with convergence rate $O(1/n)$

Participants: Eric Moulines, Francis Bach [correspondent].

Large-scale machine learning problems are becoming ubiquitous in many areas of science and engineering. Faced with large amounts of data, practitioners typically prefer algorithms that process each observation only once, or a few times. Stochastic approximation algorithms such as stochastic gradient descent (SGD) and its variants, although introduced more than sixty years ago, still remain the most widely used and studied method in this context. In [8], we consider the stochastic approximation problem where a convex function has to be minimized, given only the knowledge of unbiased estimates of its gradients at certain points, a framework which includes machine learning methods based on the minimization of the empirical risk. We focus on problems without strong convexity, for which all previously known algorithms achieve a convergence rate for function values of $O(1/\sqrt{n})$ after n iterations. We consider and analyze two algorithms that achieve a rate of $O(1/n)$ for classical supervised learning problems. For least-squares regression, we show that *averaged* stochastic gradient descent *with constant step-size* achieves the desired rate. For logistic regression, this is achieved by a simple novel stochastic gradient algorithm that (a) constructs successive local quadratic approximations of the loss functions, while (b) preserving the same running-time complexity as stochastic gradient descent. For these algorithms, we provide a non-asymptotic analysis of the generalization error (in expectation, and also in high probability for least-squares), and run extensive experiments showing that they often outperform existing approaches.

6.5. Streaming Bayesian Inference

Participant: Michael Jordan [correspondent].

Large, streaming data sets are increasingly the norm in science and technology. Simple descriptive statistics can often be readily computed with a constant number of operations for each data point in the streaming setting, without the need to revisit past data or have advance knowledge of future data. But these time and memory restrictions are not generally available for the complex, hierarchical models that practitioners often have in mind when they collect large data sets. Significant progress on scalable learning procedures has been made in recent years. But the underlying models remain simple, and the inferential framework is generally non-Bayesian. The advantages of the Bayesian paradigm (e.g., hierarchical modeling, coherent treatment of uncertainty) currently seem out of reach in the Big Data setting.

An exception to this statement is provided by Hofmann et al. (2010), who have shown that a class of approximation methods known as *variational Bayes* (VB) can be usefully deployed for large-scale data sets. They have applied their approach, referred to as *stochastic variational inference* (SVI), to the domain of topic modeling of document collections, an area with a major need for scalable inference algorithms. VB traditionally uses the variational lower bound on the marginal likelihood as an objective function, and the idea of SVI is to apply a variant of stochastic gradient descent to this objective. Notably, this objective is based on the conceptual existence of a full data set involving D data points (i.e., documents in the topic model setting), for a fixed value of D. Although the stochastic gradient is computed for a single, small subset of data points (documents) at a time, the posterior being targeted is a posterior for D data points. This value of D must be specified in advance and is used by the algorithm at each step. Posteriors for D' data points, for D' not equal to D, are not obtained as part of the analysis.

We view this lack of a link between the number of documents that have been processed thus far and the posterior that is being targeted as undesirable in many settings involving streaming data. In this project we aim at an approximate Bayesian inference algorithm that is scalable like SVI but is also truly a streaming procedure, in that it yields an approximate posterior for each processed collection of D' data points—and not just a pre-specified "final" number of data points D. To that end, we return to the classical perspective of Bayesian updating, where the recursive application of Bayes theorem provides a sequence of posteriors, not a sequence of approximations to a fixed posterior. To this classical recursive perspective we bring the VB framework; our updates need not be exact Bayesian updates but rather may be approximations such as VB.

Although the empirical success of SVI is the main motivation for our work, we are also motivated by recent developments in computer architectures, which permit distributed and asynchronous computations in addition to streaming computations. A streaming VB algorithm naturally lends itself to distributed and asynchronous implementations.

6.6. Convex Relaxations for Permutation Problems

Participants: Fajwel Fogel [correspondent], Rodolphe Jenatton, Francis Bach, Alexandre d'Aspremont.

Seriation seeks to reconstruct a linear order between variables using unsorted similarity information. It has direct applications in archeology and shotgun gene sequencing for example. In [12] we prove the equivalence between the seriation and the combinatorial 2-sum problem (a quadratic minimization problem over permutations) over a class of similarity matrices. The seriation problem can be solved exactly by a spectral algorithm in the noiseless case and we produce a convex relaxation for the 2-sum problem to improve the robustness of solutions in a noisy setting. This relaxation also allows us to impose additional structural constraints on the solution, to solve semi-supervised seriation problems. We performed numerical experiments on archeological data, Markov chains and gene sequences.

6.7. Phase retrieval for imaging problems

Participants: Fajwel Fogel [correspondent], Irène Waldspurger, Alexandre d'Aspremont.

In [29] we study convex relaxation algorithms for phase retrieval on imaging problems. We show that structural assumptions on the signal and the observations, such as sparsity, smoothness or positivity, can be exploited to both speed-up convergence and improve recovery performance. We detail experimental results in molecular imaging problems simulated from PDB data.

Phase retrieval seeks to reconstruct a complex signal, given a number of observations on the *magnitude* of linear measurements, i.e. solve

$$\begin{array}{ll} \text{find} & x \\ \text{such that} & |Ax| = b \end{array}$$

in the variable x, where A and b. This problem has direct applications in X-ray and crystallography imaging, diffraction imaging, Fourier optics or microscopy for example, in problems where physical limitations mean detectors usually capture the intensity of observations but cannot recover their phase. In this project, we focus on problems arising in diffraction imaging, where A is usually a Fourier transform, often composed with one or multiple masks (a technique sometimes called ptychography). The Fourier structure, through the FFT, often considerably speeds up basic linear operations, which allows us to solve large scale convex relaxations on realistically large imaging problems. We also observe that in most of the imaging problems we consider, the Fourier transform is very sparse, with known support (we lose the phase but observe the magnitude of Fourier coefficients), which allows us to considerably reduce the size of our convex phase retrieval relaxations.

6.8. Learning Sparse Penalties for Change-point Detection using Max Margin Interval Regression

Participants: Toby Hocking, Guillem Rigaill, Jean-Philippe Vert, Francis Bach [correspondent].

In segmentation models, the number of change-points is typically chosen using a penalized cost function. In [22] we propose to learn the penalty and its constants in databases of signals with weak change-point annotations. We propose a convex relaxation for the resulting interval regression problem, and solve it using accelerated proximal gradient methods. We show that this method achieves state-of-the-art change-point detection in a database of annotated DNA copy number profiles from neuroblastoma tumors.

6.9. Maximizing submodular functions using probabilistic graphical models

Participants: K. S. Sesh Kumar [correspondent], Francis Bach.

In [34] we consider the problem of maximizing submodular functions; while this problem is known to be NP-hard, several numerically efficient local search techniques with approximation guarantees are available. In this paper, we propose a novel convex relaxation which is based on the relationship between submodular functions, entropies and probabilistic graphical models. In a graphical model, the entropy of the joint distribution decomposes as a sum of marginal entropies of subsets of variables; moreover, for any distribution, the entropy of the closest distribution factorizing in the graphical model provides an bound on the entropy. For directed graphical models, this last property turns out to be a direct consequence of the submodularity of the entropy function, and allows the generalization of graphical-model-based upper bounds to any submodular functions. These upper bounds may then be jointly maximized with respect to a set, while minimized with respect to the graph, leading to a convex variational inference scheme for maximizing submodular functions, based on outer approximations of the marginal polytope and maximum likelihood bounded treewidth structures. By considering graphs of increasing treewidths, we may then explore the trade-off between computational complexity and tightness of the relaxation. We also present extensions to constrained problems and maximizing the difference of submodular functions, which include all possible set functions.

Optimizing submodular functions has been an active area of research with applications in graph-cut-based image segmentation [44], sensor placement [69], or document summarization [70]. A set function F is a function defined on the power set 2^V of a certain set V. It is submodular if and only if for all $A, B \subseteq V$, $F(A) + F(B) \geqslant F(A \cap B) + F(A \cup B)$. Equivalently, these functions also admit the diminishing returns property, i.e., the marginal cost of an element in the context of a smaller set is more than its cost in the context of a larger set. Classical examples of such functions are entropy, mutual information, cut functions, and covering functions—see further examples in [58], [38].

Submodular functions form an interesting class of discrete functions because minimizing a submodular function can be done in polynomial time [58], while maximization, although NP-hard, admits constant factor approximation algorithms [76]. In this paper, our ultimate goal is to provide the first (to the best of our knowledge) generic convex relaxation of submodular function maximization, with a hierarchy of complexities related to known combinatorial hierarchies such as the Sherali-Adams hierarchy [83]. Beyond the graphical model tools that we are going to develop, having convex relaxations may be interesting for several reasons: (1) they can lead to better solutions, (2) they provide online bounds that may be used within branch-and-bound optimization and (3) they ease the use of such combinatorial optimization problems within structured prediction framework [91].

We make the following contributions:

- For any directed acyclic graph G and a submodular function F, we define a bound $F_G(A)$ and study its properties (monotonicity, tightness), which is specialized to decomposable graphs.
- We propose an algorithm to maximize submodular functions by maximizing the bound $F_G(A)$ with respect to A while minimizing with respect to the graph G, leading to a convex variational method based on outer approximation of the marginal polytope [93] and inner approximation of the hypertree polytope.
- We propose extensions to constrained problems and maximizing the difference of submodular functions, which include all possible set functions.
- We illustrate our results on small-scale experiments.

6.10. Reflection methods for user-friendly submodular optimization

Participants: Stefanie Jegelka, Suvrit Sra, Francis Bach [correspondent].

Recently, it has become evident that submodularity naturally captures widely occurring concepts in machine learning, signal processing and computer vision. Consequently, there is need for efficient optimization procedures for submodular functions, especially for minimization problems. While general submodular minimization is challenging, we propose in [15] a new method that exploits existing decomposability of submodular functions. In contrast to previous approaches, our method is neither approximate, nor impractical, nor does it need any cumbersome parameter tuning. Moreover, it is easy to implement and parallelize. A key

component of our method is a formulation of the discrete submodular minimization problem as a continuous best approximation problem that is solved through a sequence of reflections, and its solution can be easily thresholded to obtain an optimal discrete solution. This method solves *both* the continuous and discrete formulations of the problem, and therefore has applications in learning, inference, and reconstruction. In our experiments, we illustrate the benefits of our method on two image segmentation tasks.

6.11. Convex Relaxations for Learning Bounded Treewidth Decomposable Graphs

Participants: K. S. Sesh Kumar [correspondent], Francis Bach.

In [24] we consider the problem of learning the structure of undirected graphical models with bounded treewidth, within the maximum likelihood framework. This is an NP-hard problem and most approaches consider local search techniques. In this paper, we pose it as a combinatorial optimization problem, which is then relaxed to a convex optimization problem that involves searching over the forest and hyperforest polytopes with special structures, independently. A supergradient method is used to solve the dual problem, with a run-time complexity of $O(k^3 n^{k+2} \log n)$ for each iteration, where n is the number of variables and k is a bound on the treewidth. We compare our approach to state-of-the-art methods on synthetic datasets and classical benchmarks, showing the gains of the novel convex approach.

Graphical models provide a versatile set of tools for probabilistic modeling of large collections of interdependent variables. They are defined by graphs that encode the conditional independences among the random variables, together with potential functions or conditional probability distributions that encode the specific local interactions leading to globally well-defined probability distributions [42], [93], [67].

In many domains such as computer vision, natural language processing or bioinformatics, the structure of the graph follows naturally from the constraints of the problem at hand. In other situations, it might be desirable to estimate this structure from a set of observations. It allows (a) a statistical fit of rich probability distributions that can be considered for further use, and (b) discovery of structural relationship between different variables. In the former case, distributions with tractable inference are often desirable, i.e., inference with run-time complexity does not scale exponentially in the number of variables in the model. The simplest constraint to ensure tractability is to impose tree-structured graphs [52]. However, these distributions are not rich enough, and following earlier work [73], [39], [75], [48], [59], [89], we consider models with *treewidth* bounded, not simply by one (i.e., trees), but by a small constant k.

Beyond the possibility of fitting tractable distributions (for which probabilistic inference has linear complexity in the number of variables), learning bounded-treewidth graphical models is key to design approximate inference algorithms for graphs with higher treewidth. Indeed, as shown by [82], [93], [68], approximating general distributions by tractable distributions is a common tool in variational inference. However, in practice, the complexity of variational distributions is often limited to trees (i.e., $k = 1$), since these are the only ones with exact polynomial-time structure learning algorithms. The convex relaxation we designed enables us to augment the applicability of variational inference, by allowing a finer trade-off between run-time complexity and approximation quality.

We make the following contributions:

- We provide a novel convex relaxation for learning bounded-treewidth decomposable graphical models from data in polynomial time. This is achieved by posing the problem as a combinatorial optimization problem, which is relaxed to a convex optimization problem that involves the graphic and hypergraphic matroids.
- We show how a supergradient ascent method may be used to solve the dual optimization problem, using greedy algorithms as inner loops on the two matroids. Each iteration has a run-time complexity of $O(k^3 n^{k+2} \log n)$, where n is the number of variables. We also show how to round the obtained fractional solution.
- We compare our approach to state-of-the-art methods on synthetic datasets and classical benchmarks showing the gains of the novel convex approach.

6.12. Large-Margin Metric Learning for Partitioning Problems

Participants: Rémi Lajugie [correspondent], Sylvain Arlot, Francis Bach.

In [31] we consider unsupervised partitioning problems, such as clustering, image segmentation, video segmentation and other change-point detection problems. We focus on partitioning problems based explicitly or implicitly on the minimization of Euclidean distortions, which include mean-based change-point detection, K-means, spectral clustering and normalized cuts. Our main goal is to learn a Mahalanobis metric for these unsupervised problems, leading to feature weighting and/or selection. This is done in a supervised way by assuming the availability of several potentially partially labelled datasets that share the same metric. We cast the metric learning problem as a large-margin structured prediction problem, with proper definition of regularizers and losses, leading to a convex optimization problem which can be solved efficiently with iterative techniques. We provide experiments where we show how learning the metric may significantly improve the partitioning performance in synthetic examples, bioinformatics, video segmentation and image segmentation problems.

Unsupervised partitioning problems are ubiquitous in machine learning and other data-oriented fields such as computer vision, bioinformatics or signal processing. They include (a) traditional *unsupervised clustering* problems, with the classical K-means algorithm, hierarchical linkage methods [61] and spectral clustering [80], (b) *unsupervised image segmentation* problems where two neighboring pixels are encouraged to be in the same cluster, with mean-shift techniques [51] or normalized cuts [84], and (c) *change-point detection* problems adapted to multivariate sequences (such as video) where segments are composed of contiguous elements, with typical window-based algorithms [54] and various methods looking for a change in the mean of the features (see, e.g., [49]).

All the algorithms mentioned above rely on a specific distance (or more generally a similarity measure) on the space of configurations. A good metric is crucial to the performance of these partitioning algorithms and its choice is heavily problem-dependent. While the choice of such a metric has been originally tackled manually (often by trial and error), recent work has considered learning such metric directly from data. Without any supervision, the problem is ill-posed and methods based on generative models may learn a metric or reduce dimensionality (see, e.g., [53]), but typically with no guarantees that they lead to better partitions. In this paper, we follow [41], [95], [40] and consider the goal of learning a metric for potentially several partitioning problems sharing the same metric, assuming that several fully or partially labelled partitioned datasets are available during the learning phase. While such labelled datasets are typically expensive to produce, there are several scenarios where these datasets have already been built, often for evaluation purposes. These occur in video segmentation tasks, image segmentation tasks as well as change-point detection tasks in bioinformatics (see [62]).

W consider partitioning problems based explicitly or implicitly on the minimization of Euclidean distortions, which include K-means, spectral clustering and normalized cuts, and mean-based change-point detection. We make the following contributions:

- We review and unify several partitioning algorithms, and cast them as the maximization of a linear function of a rescaled equivalence matrix, which can be solved by algorithms based on spectral relaxations or dynamic programming.
- Given fully labelled datasets, we cast the metric learning problem as a large-margin structured prediction problem, with proper definition of regularizers, losses and efficient loss-augmented inference.
- Given partially labelled datasets, we propose an algorithm, iterating between labeling the full datasets given a metric and learning a metric given the fully labelled datasets. We also consider extensions that allow changes in the full distribution of univariate time series (rather than changes only in the mean), with application to bioinformatics.
- We provide experiments where we show how learning the metric may significantly improve the partitioning performance in synthetic examples, video segmentation and image segmentation problems.

6.13. Comparison between multi-task and single-task oracle risks in kernel ridge regression

Participant: Matthieu Solnon [correspondent].

In [35] we study multi-task kernel ridge regression and try to understand when the multi-task procedure performs better than the single-task one, in terms of averaged quadratic risk. In order to do so, we compare the risks of the estimators with perfect calibration, the oracle risk. We are able to give explicit settings, favorable to the multi-task procedure, where the multi-task oracle performs better than the single-task one. In situations where the multi-task procedure is conjectured to perform badly, we also show the oracle does so. We then complete our study with simulated examples, where we can compare both oracle risks in more natural situations. A consequence of our result is that the multi-task ridge estimator has a lower risk than any single-task estimator, in favorable situations.

Increasing the sample size is the most common way to improve the performance of statistical estimators. In some cases (see, for instance, the experiments of [56] on customer data analysis or those of [63] on molecule binding problems), having access to some new data may be impossible, often due to experimental limitations. One way to circumvent those constraints is to use datasets from several related (and, hopefully, "similar") problems, as if it gave additional (in some sense) observations on the initial problem. The statistical methods using this heuristic are called "multi-task" techniques, as opposed to "single-task" techniques, where every problem is treated one at a time. In this paper, we study kernel ridge regression in a multi-task framework and try to understand when multi-task can improve over single-task.

The first trace of a multi-task estimator can be found in the work of [88]. In this article, Charles Stein showed that the usual maximum-likelihood estimator of the mean of a Gaussian vector (of dimension larger than 3, every dimension representing here a task) is not admissible—that is, there exists another estimator that has a lower risk for every parameter. He showed the existence of an estimator that uniformly attains a lower quadratic risk by shrinking the estimators along the different dimensions towards an arbitrary point. An explicit form of such an estimator was given by [64], yielding the famous James-Stein estimator. This phenomenon, now known as the "Stein's paradox", was widely studied in the following years and the behaviour of this estimator was confirmed by empirical studies, in particular the one from [55]. This first example clearly shows the goals of the multi-task procedure: an advantage is gained by borrowing information from different tasks (here, by shrinking the estimators along the different dimensions towards a common point), the improvement being scored by the global (averaged) squared risk. Therefore, this procedure does not guarantee individual gains on every task, but a global improvement on the sum of those task-wise risks.

We consider $p \geq 2$ different regression tasks, a framework we refer to as "multi-task" regression, and where the performance of the estimators is measured by the fixed-design quadratic risk. Kernel ridge regression is a classical framework to work with and comes with a natural norm, which often has desirable properties (such as, for instance, links with regularity). This norm is also a natural "similarity measure" between the regression functions. [56] showed how to extend kernel ridge regression to a multi-task setting, by adding a regularization term that binds the regression functions along the different tasks together. One of the main questions that is asked is to assert whether the multi-task estimator has a lower risk than any single-task estimator. It was recently proved by [86] that a fully data-driven calibration of this procedure is possible, given some assumptions on the set of matrices used to regularize—which correspond to prior knowledge on the tasks. Under those assumptions, the estimator is showed to verify an *oracle inequality*, that is, its risk matches (up to constants) the best possible one, the *oracle risk*. Thus, it suffices to compare the oracle risks for the multi-task procedure and the single-task one to provide an answer to this question.

We study the oracle multi-task risk and compare it to the oracle single-task risk. We then find situations where the multi-task oracle is proved to have a lower risk than the single-task oracle. This allows us to better understand which situation favors the multi-task procedure and which does not. After having defined our model, we write down the risk of a general multi-task ridge estimator and see that it admits a convenient decomposition using two key elements: the mean of the tasks and the resulting variance. This decomposition allows us to optimize this risk and get a precise estimation of the oracle risk, in settings where the ridge

estimator is known to be minimax optimal. We then explore several repartitions of the tasks that give the latter multi-task rates, study their single-task oracle risk and compare it to their respective multi-task rates. This allows us to discriminate several situations, depending whether the multi-task oracle either outperforms its single-task counterpart, underperforms it or whether both behave similarly. We also show that, in the cases favorable to the multi-task oracle detailed in the previous sections, the estimator proposed by [86] behaves accordingly and achieves a lower risk than the single-task oracle. We finally study settings where we can no longer explicitly study the oracle risk, by running simulations, and we show that the multi-task oracle continues to retain the same virtues and disadvantages as before.

6.14. Sharp analysis of low-rank kernel matrix approximations

Participant: Francis Bach [correspondent].

Kernel methods, such as the support vector machine or kernel ridge regression, are now widely used in many areas of science and engineering, such as computer vision or bioinformatics. However, kernel methods typically suffer from at least quadratic running-time complexity in the number of observations n, as this is the complexity of computing the kernel matrix. In large-scale settings where n may be large, this is usually not acceptable. In [7], we consider supervised learning problems within the positive-definite kernel framework, such as kernel ridge regression, kernel logistic regression or the support vector machine. Low-rank approximations of the kernel matrix are often considered as they allow the reduction of running time complexities to $O(p^2 n)$, where p is the rank of the approximation. The practicality of such methods thus depends on the required rank p. In this paper, we show that in the context of kernel ridge regression, for approximations based on a random subset of columns of the original kernel matrix, the rank p may be chosen to be linear in the *degrees of freedom* associated with the problem, a quantity which is classically used in the statistical analysis of such methods, and is often seen as the implicit number of parameters of non-parametric estimators. This result enables simple algorithms that have sub-quadratic running time complexity, but provably exhibit the same *predictive performance* than existing algorithms, for any given problem instance, and not only for worst-case situations.

6.15. fMRI encoding and decoding models

Participant: Fabian Pedregosa [correspondent].

In [20] we show that HRF estimation improves sensitivity of fMRI encoding and decoding models and propose a new approach for the estimation of Hemodynamic Response Functions from fMRI data. The model we propose is based on the linearity assumption behind the General Linear Model and can be computed using standard gradient-based solvers. We use the activation patterns computed by our model as input data for encoding and decoding studies and report performance improvement in both settings.

This work proves that significant improvements in recovery of brain activation patterns can be made by estimating the form of the Hemodynamic Response Function instead of using a canonical form for this response.

6.16. Structured Penalties for Log-linear Language Models

Participants: Anil Nelakanti [correspondent], Cédric Archambeau, Francis Bach, Guillaume Bouchard.

Language models can be formalized as log-linear regression models where the input features represent previously observed contexts up to a certain length m. The complexity of existing algorithms to learn the parameters by maximum likelihood scale linearly in nd, where n is the length of the training corpus and d is the number of observed features. In [19] we present a model that grows logarithmically in d, making it possible to efficiently leverage longer contexts. We account for the sequential structure of natural language using tree-structured penalized objectives to avoid overfitting and achieve better generalization.

Language models are crucial parts of advanced natural language processing pipelines, such as speech recognition [45], machine translation [47], or information retrieval [92]. When a sequence of symbols is observed, a language model predicts the probability of occurrence of the next symbol in the sequence. Models based on so-called back-off smoothing have shown good predictive power [60]. In particular, Kneser-Ney (KN) and its variants [66] are still achieving state-of-the-art results for more than a decade after they were originally proposed. Smoothing methods are in fact clever heuristics that require tuning parameters in an ad-hoc fashion. Hence, more principled ways of learning language models have been proposed based on maximum entropy [50] or conditional random fields [81], or by adopting a Bayesian approach [94].

We focus on penalized maximum likelihood estimation in log-linear models. In contrast to language models based on *unstructured* norms such as ℓ_2 (quadratic penalties) or ℓ_1 (absolute discounting), we use *tree-structured* norms [96], [65]. Structured penalties have been successfully applied to various NLP tasks, including chunking and named entity recognition [74], but not language modeling. Such penalties are particularly well-suited to this problem as they mimic the nested nature of word contexts. However, existing optimizing techniques are not scalable for large contexts m.

We show that structured tree norms provide an efficient framework for language modeling. Furthermore, we give the first algorithm for structured ℓ_∞ tree norms with a complexity nearly linear in the number of nodes. This leads to a memory-efficient *and* time-efficient learning algorithm for generalized linear language models.

6.17. Distributed Large-scale Natural Graph Factorization

Participants: Amr Ahmed, Nino Shervashidze [correspondent], Shravan Narayanamurthy, Vanja Josifovski, Alexander Smola.

Natural graphs, such as social networks, email graphs, or instant messaging patterns, have become pervasive through the internet. These graphs are massive, often containing hundreds of millions of nodes and billions of edges. While some theoretical models have been proposed to study such graphs, their analysis is still difficult due to the scale and nature of the data. We propose a framework for large-scale graph decomposition and inference. To resolve the scale, our framework in [6] is distributed so that the data are partitioned over a shared-nothing set of machines. We propose a novel factorization technique that relies on partitioning a graph so as to minimize the number of neighboring vertices rather than edges across partitions. Our decomposition is based on a streaming algorithm. It is network-aware as it adapts to the network topology of the underlying computational hardware. We use local copies of the variables and an efficient asynchronous communication protocol to synchronize the replicated values in order to perform most of the computation without having to incur the cost of network communication. On a graph of 200 million vertices and 10 billion edges, derived from an email communication network, our algorithm retains convergence properties while allowing for almost linear scalability in the number of computers.

6.18. Evaluating Speech Features with the Minimal-Pair ABX task

Participants: Thomas Schatz [correspondent], Vijayaditya Peddinti, Francis Bach, Aren Jansen, Hynek Hermansky, Emmanuel Dupoux.

In [23] we introduce a new framework for the evaluation of speech representations in zero-resource settings, that extends and complements previous work by Carlin, Jansen and Hermansky [46]. In particular, we replace their Same/Different discrimination task by several Minimal-Pair ABX (MP-ABX) tasks. We explain the analytical advantages of this new framework and apply it to decompose the standard signal processing pipelines for computing PLP and MFC coefficients. This method enables us to confirm and quantify a variety of well-known and not-so-well-known results in a single framework.

Speech recognition technology crucially rests on adequate speech features for encoding input data. Several such features have been proposed and studied (MFCCs, PLPs, etc), but they are often evaluated indirectly using complex tasks like phone classification or word identification. Such an evaluation technique suffers from several limitations. First, it requires a large enough annotated corpus in order to train the classifier/recognizer. Such a resource may not be available in all languages or dialects (the so-called "zero or limited resource" setting). Second, supervised classifiers may be too powerful and may compensate for potential defects of speech features (for instance noisy/unreliable channels). However, such defects are problematic in unsupervised learning techniques. Finally, the particular statistical assumptions of the classifier (linear, Gaussian, etc.) may not be suited for specific speech features (for instance sparse neural codes as in Hermansky [85]). It is therefore important to replace these complex evaluation schemes by simpler ones which tap more directly the properties of the speech features.

We extend and complement the framework proposed by Carlin, Jansen and Hermansky [46] for the evaluation of speech features in zero resource settings. This framework uses a Same-Different word discrimination task that does not depend on phonetically labelled data, nor on training a classifier. It assumes a speech corpus segmented into words, and derives a word-by-word acoustic distance matrix computed by comparing every word with every other one using Dynamic Time Warping (DTW). Carlin et al. then compute an average precision score which is used to evaluate speech features (the higher average precision, the better the features).

We explore an extension of this technique through Minimal-Pair ABX tasks (MP-ABX tasks) tested on a phonetically balanced corpus [57]. This improves the interpretability of the Carlin et al evaluation results in three different ways. First, the Same/Different task requires the computation of a ROC curve in order to derive average precision. In contrast, the ABX task is a discrimination task used in psychophysics (see [72], chapter 9) which allows for the direct computation of an error rate or a d' measure that are easier to interpret than average precision [46] and involve no assumption about ROC curves. Second, the Same/Different task compares *sets of words*, and as a result is influenced by the mix of similar versus distinct words or short versus long words in the corpus. The ABX task, in contrast, is computed on *word pairs*, and therefore enables to make linguistically precise comparisons, as in word *minimal pairs*, i.e. words differing by only one phoneme. Variants of the task enable to study phoneme discrimination across talkers and/or phonetic contexts, as well as talker discrimination across phonemes. Because it is more controlled and provides a parameter and model-free metric, the MP-ABX error rate also enables to compare performance across databases or across languages. Third, we compute bootstrap-based estimates of the variability of our performance measures, which allows us to derive confidence intervals for the error rates and tests of the significance of the difference between the error rates obtained with different representations.

6.19. Hidden Markov Tree Models for Semantic Class Induction

Participants: Edouard Grave [correspondent], Guillaume Obozinski, Francis Bach.

In [13] we propose a new method for semantic class induction. First, we introduce a generative model of sentences, based on dependency trees and which takes into account homonymy. Our model can thus be seen as a generalization of Brown clustering. Second, we describe an efficient algorithm to perform inference and learning in this model. Third, we apply our proposed method on two large datasets (10^8 tokens, 10^5 words types), and demonstrate that classes induced by our algorithm improve performance over Brown clustering on the task of semi-supervised supersense tagging and named entity recognition.

Most competitive learning methods for computational linguistics are supervised, and thus require labeled examples, which are expensive to obtain. Moreover, those techniques suffer from data scarcity: many words only appear a small number of time, or even not at all, in the training data. It thus helps a lot to first learn word clusters on a large amount of unlabeled data, which are cheap to obtain, and then to use this clusters as features for the supervised task. This scheme has proven to be effective for various tasks such as named entity recognition, syntactic chunking or syntactic dependency parsing. It was also successfully applied for transfer learning of multilingual structure.

The most commonly used clustering method for semi-supervised learning is known as Brown clustering. While still being one of the most efficient word representation method, Brown clustering has two limitations we want to address in this work. First, since it is a hard clustering method, homonymy is ignored. Second, it does not take into account syntactic relations between words, which seems crucial to induce semantic classes. Our goal is thus to propose a method for semantic class induction which takes into account both syntax and homonymy, and then to study their effects on semantic class learning.

We start by introducing a new unsupervised method for semantic classes induction. This is achieved by defining a generative model of sentences with latent variables, which aims at capturing semantic roles of words. We require our method to be scalable, in order to learn models on large datasets containing tens of millions of sentences. More precisely, we make the following contributions:

- We introduce a generative model of sentences, based on dependency trees, which can be seen as a generalization of Brown clustering,
- We describe a fast approximate inference algorithm, based on message passing and online EM for scaling to large datasets. It allowed us to learn models with 512 latent states on a dataset with hundreds of millions of tokens in less than two days on a single core,
- We learn models on two datasets, Wikipedia articles about musicians and the NYT corpus, and evaluate them on two semi-supervised tasks, namely supersense tagging and named entity recognition.

6.20. Domain Adaptation for Sequence Labeling using Hidden Markov Models

Participants: Edouard Grave [correspondent], Guillaume Obozinski, Francis Bach.

Most natural language processing systems based on machine learning are not robust to domain shift. For example, a state-of-the-art syntactic dependency parser trained on Wall Street Journal sentences has an absolute drop in performance of more than ten points when tested on textual data from the Web. An efficient solution to make these methods more robust to domain shift is to first learn a word representation using large amounts of unlabeled data from both domains, and then use this representation as features in a supervised learning algorithm. In this paper, we propose to use hidden Markov models to learn word representations for part-of-speech tagging. In particular, we study the influence of using data from the source, the target or both domains to learn the representation and the different ways to represent words using an HMM.

Nowadays, most natural language processing systems are based on supervised machine learning. Despite the great successes obtained by those techniques, they unfortunately still suffer from important limitations. One of them is their sensitivity to domain shift: for example, a state-of-the-art part-of-speech tagger trained on the Wall Street Journal section of the Penn treebank achieves an accuracy of 97% when tested on sentences from the Wall Street Journal, but only 90% when tested on textual data from the Web. This drop in performance can also be observed for other tasks such as syntactic parsing or named entity recognition.

One of the explanations for this drop in performance is the big lexical difference that exists accross domains. This results in a lot of out-of-vocabulary words (OOV) in the test data, *i.e.*, words of the test data that were not observed in the training set. For example, more than 25% of the tokens of the test data from the Web corpus are unobserved in the training data from the WSJ. By comparison, only 11.5% of the tokens of the test data from the WSJ are unobserved in the training data from the WSJ. Part-of-speech taggers make most of their errors on those out-of-vocabulary words.

Labeling enough data to obtain a high accuracy for each new domain is not a viable solution. Indeed, it is expensive to label data for natural language processing, because it requires expert knowledge in linguistics. Thus, there is an important need for transfer learning, and more precisely for domain adaptation, in computational linguistics. A common solution consists in using large quantities of unlabeled data, from both source and target domains, in order to learn a good word representation. This representation is then used as features to train a supervised classifier that is more robust to domain shift. Depending on how much data from the source and the target domains are used, this method can be viewed as performing semi-supervised learning or domain adaptation. The goal is to reduce the impact of out-of-vocabulary words on performance. This scheme was first proposed to reduce data sparsity for named entity recognition, before being applied to domain adaptation for part-of-speech tagging or syntactic parsing.

Hidden Markov models have already been considered in previous work to learn word representations for domain adaptation or semi-supervised learning. Our contributions in [25] are mostly experimental: we compare different word representations that can be obtained from an HMM and study the effect of training the unsupervised HMM on source, target or both domains. While previous work mostly use Viterbi decoding to obtain word representations from an HMM, we empirically show that posterior distributions over latent classes give better results.

6.21. Simple Greedy Matching for Aligning Large Knowledge Bases

Participant: Simon Lacoste-Julien [correspondent].

Collaboration with: Konstantina Palla, Alex Davies, Zoubin Ghahramani (Machine Learning Group, Department of Engineering, University of Cambridge), Gjergji Kasneci (Max Planck Institut für Informatik), Thore Graepel (Microsoft Research Cambridge)

The Internet has enabled the creation of a growing number of large-scale knowledge bases in a variety of domains containing complementary information. Tools for automatically aligning these knowledge bases would make it possible to unify many sources of structured knowledge and answer complex queries. However, the efficient alignment of large-scale knowledge bases still poses a considerable challenge. Here, we present Simple Greedy Matching (SiGMa), a simple algorithm for aligning knowledge bases with millions of entities and facts. SiGMa is an iterative propagation algorithm that leverages both the structural information from the relationship graph and flexible similarity measures between entity properties in a greedy local search, which makes it scalable. Despite its greedy nature, our experiments in [17] indicate that SiGMa can efficiently match some of the world's largest knowledge bases with high accuracy. We provide additional experiments on benchmark datasets which demonstrate that SiGMa can outperform state-of-the-art approaches both in accuracy and efficiency.

7. Bilateral Contracts and Grants with Industry

7.1. Bilateral Contracts with Industry

- Technicolor: "Tensor factorization algorithms for recommendation systems".
- Xerox: CIFRE PhD student "IMAGE2TXT: From images to text".
- 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 challeges 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-Rocquencourt and Grenoble) and four MSR sites (Cambridge, New England, Redmond, New York).

7.2. Bilateral Grants with Industry

- Google Research Award: "Large scale adaptive machine learning with finite data sets"

8. Partnerships and Cooperations

8.1. National Initiatives

8.1.1. ANR: Calibration

Participant: Sylvain Arlot.

S. Arlot, Membre du projet ANR Calibration

Titre: Statistical calibration

Coordinator: University Paris Dauphine

Leader: Vincent Rivoirard

Other members: 34 members, mostly among CEREMADE (Paris Dauphine), Laboratoire Jean-Alexandre Dieudonné (Université de Nice) and Laboratoire de Mathématiques de l'Université Paris Sud

Instrument: ANR Blanc

Duration: Jan 2012 - Dec 2015

Total funding: 240 000 euros

Webpage: https://sites.google.com/site/anrcalibration/

8.1.2. CNRS: Gargantua

Participants: Sylvain Arlot, Francis Bach.

S. Arlot, F. Bach, membres du projet "Gargantua"

Titre: Big data; apprentissage automatique et optimisation mathématique pour les données gigantesques

Coordinator: Laboratoire Jean Kuntzmann (UMR 5224)

Leader: Zaid Harchaoui

Other members: 13 members: S. Arlot, F. Bach and researchers from Laboratoire Jean Kuntzmann, Laboratoire d'Informatique de Grenoble (Universite Joseph Fourier) and Laboratoire Paul Painleve (Universite Lille 1).

Instrument: défi MASTODONS du CNRS

Duration: May 2013-Dec 2013 (may be reconducted for 2014)

Total funding: 30 000 euros for 2013

Webpage: http://lear.inrialpes.fr/people/harchaoui/projects/gargantua/index.html

8.2. European Initiatives

8.2.1. SIERRA

Participants: Francis Bach [correspondent], Simon Lacoste-Julien, Augustin Lefèvre, Nicolas Le Roux, Mark Schmidt.

Type: IDEAS

Instrument: ERC Starting Grant

Duration: December 2009 - November 2014

Coordinator: Inria (France)

Abstract: Machine learning is now a core part of many research domains, where the abundance of data has forced researchers to rely on automated processing of information. The main current paradigm of application of machine learning techniques consists in two sequential stages: in the representation phase, practitioners first build a large set of features and potential responses for model building or prediction. Then, in the learning phase, off-the-shelf algorithms are used to solve the appropriate data processing tasks. While this has led to significant advances in many domains, the potential of machine learning techniques is far from being reached.

8.2.2. SIPA

Participants: Alexandre d'Aspremont [correspondent], Fajwel Fogel.

Type: IDEAS
Instrument: ERC Starting Grant
Duration: May 2011 - May 2016
Coordinator: 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.

8.3. International Initiatives

8.3.1. Inria Associate Team STATWEB

Title: Fast Statistical Analysis of Web Data via Sparse Learning
Inria principal investigator: Francis Bach
International Partner (Institution - Laboratory - Researcher):

University of California Berkeley (United States) - EECS and IEOR Departments - Francis Bach

Duration: 2011 - 2013
See also: http://www.di.ens.fr/~fbach/statweb.html
The goal of the proposed research is to provide web-based tools for the analysis and visualization of large corpora of text documents, with a focus on databases of news articles. We intend to use advanced algorithms, drawing from recent progresses in machine learning and statistics, to allow a user to quickly produce a short summary and associated timeline showing how a certain topic is described in news media. We are also interested in unsupervised learning techniques that allow a user to understand the difference between several different news sources, topics or documents.

8.4. International Research Visitors

8.4.1. Visits of International Scientists

Michael Jordan (U.C. Berkeley), spent one year in our team, until the summer 2013, financed by the Fondation de Sciences Mathématiques de Paris and Inria.

9. Dissemination

9.1. Scientific Animation

9.1.1. Editorial boards

- F. Bach: Journal of Machine Learning Research, Action Editor.
- F. Bach: IEEE Transactions on Pattern Analysis and Machine Intelligence, Associate Editor.
- F. Bach: Information and Inference, Associate Editor.
- F. Bach: SIAM Journal on Imaging Sciences, Associate Editor.
- F. Bach: International Journal of Computer Vision, Associate Editor
- A. d'Aspremont: Optimization Methods and Software
- A. d'Aspremont: SIAM Journal on Optimization

9.1.2. Area chair

- F. Bach: International Conference on Machine Learning, 2013
- F. Bach: Neural Information Processing Systems, 2013

9.1.3. Workshop and conference organization

- S. Arlot, member of the program committee of the Second Workshop on Industry & Practices for Forecasting (WIPFOR), EDF R&D, Clamart. 5-7 June 2013.
- A. d'Aspremont was co-organizer of the workshop on optimization and machine learning at Les Houches in Janurary 2013.
- F. Bach organized a workshop on "Big data: theoretical and practical challenges" - May, 14-15, 2013 - Institut Poincaré (co-organized with Michael Jordan), funded by the Fondation de Sciences Mathématiques de Paris and Inria.
- F. Bach and Michael Jordan coorganized the "Fête Parisienne in Computation, Inference and Optimization: A Young Researchers' Forum". A workshop organized in the framework of the the Schlumberger Chair for mathematical sciences at IHÉS. March 20, 2013. http://www.di.ens.fr/~fbach/ihes.html
- F. Bach also coorganized the "Workshop on "Succinct Data Representations and Applications", Theoretical Foundations of Big data. Simons Institute, Berkeley, September 2013.

9.1.4. Other

- S. Arlot is member of the board for the entrance exam in Ecole Normale Supérieure (mathematics, voie B/L).
- A. d'Aspremont is a member of the scientific committee of the programme Gaspard Monge pour l'optimisation (PGMO).
- A. d'Aspremont is a member of the scientific committee of Thales Alenia Space.

9.1.5. Invited presentations

- S. Arlot, "Kernel change-point detection", Workshop "Non-stationarity in Statistics and Risk Management" (CIRM, Marseille, January, 21-25, 2013).
- S. Arlot, "Sélection de modèles par validation croisée et sélection de paramètres pour la régression ridge et le Lasso", Groupe de Travail Neurospin-Select (Saclay, February, 20, 2013).
- S. Arlot, "Optimal model selection with V-fold cross-validation: how should V be chosen?", Fête Parisienne in Computation, Inference and Optimization: A Young Researchers' Forum (IHES, Bures-sur-Yvette, March, 20, 2013).

- S. Arlot, "Kernel change-point detection", Groupe de Travail de Statistique de Jussieu (Paris, November, 11, 2013).
- S. Arlot, "Analyse du biais de forêts purement aléatoires", Séminaire de l'Equipe de Probabilités et Statistiques (Institut Elie Cartan, Nancy, November, 28, 2013).
- S. Arlot, "Optimal data-driven estimator selection with minimal penalties", keynote lecture, Workshop "Mathematical Statistics with Applications in Mind" (CIRM, Marseille, December, 9-13, 2013).
- Simon Lacoste-Julien, "Harnessing the structure of data for discriminative machine learning":
  - Department of Statistics, University of Oxford, February 2013
  - Intelligent Systems Lab Amsterdam, University of Amsterdam, February 2013
  - Département d'informatique, Université de Sherbrooke, April 2013
  - School of Computer Science, McGill University, April 2013
  - Département d'Informatique, École Normale Supérieure, April 2013
- "Block-Coordinate Frank-Wolfe Optimization for Structured SVMs"
  - ICML, Atlanta, USA, June 2013
  - ICCOPT, Lisbon, Portugal, July 2013
- Simon Lacoste-Julien, "SiGMa: Simple Greedy Matching for Aligning Large Knowledge Bases", KDD, Chicago, August 2013
- Simon Lacoste-Julien, "Frank-Wolfe optimization insights in machine learning"
  - Département d'informatique, Université de Sherbrooke, August 2013
  - SMILE seminar, Paris, November 2013
  - SAIL meeting, UC Berkeley, December 2013
  - CILVR Lab , New York University, December 2013
  - Machine Learning Lab, Columbia University, December 2013
  - Reasoning and Learning Lab, McGill University, December 2013
- Simon Lacoste-Julien, "Making Sense of Big Data", CaFFEET, Stanford University, November
- Michael Jordan, Keynote Speaker, ACM Conference on Knowledge Discovery and Data Mining (SIGKDD), Beijing, China, 8/15/12
- Michael Jordan, Keynote Speaker, 21st Century Computing Conference, Tianjin, China, 10/25/12
- Michael Jordan, Keynote Speaker, ICONIP, Doha, Qatar, 11/12/12
- Michael Jordan, Invited Speaker, SAMSI Workshop on Massive Data Analysis, 9/9/12
- Michael Jordan, Invited Speaker, Méthodes Bayésiennes non Paramétriques pour le Traitement du
- Michael Jordan, Signal et des Images, Telecom ParisTech, Paris, France, 9/8/12
- Michael Jordan, Invited Speaker, Seminaire Parisien de Statistique, Paris, France, 9/17/12
- Michael Jordan, Invited Speaker, Workshop on Random Matrices and their Applications, Paris, France, 10/9/12
- Michael Jordan, Colloquium, Department of Informatique, Ecole Normale Superieure, 10/2/12
- Michael Jordan, Vincent Meyer Colloquium, Israel Institute of Technology, 11/5/12
- Michael Jordan, Invited Speaker, Workshop on Optimization and Statistical Learning, Les Houches, France, 1/8/13
- Michael Jordan, Harry Nyquist Lecture, Department of Electrical Engineering, Yale, 1/23/13
- Michael Jordan, Invited Speaker, Simons Workshop on Big Data, New York, 1/24/13
- Michael Jordan, Keynote Speaker, Workshop on Nonsmooth Optimization in Machine Learning, Liege, Belgium, 3/4/13

- Michael Jordan, Keynote Speaker, StatLearn Workshop, Bordeaux, France, 4/8/13
- Michael Jordan, Lecture Series, Ecole Nationale de la Statistique et de l'Administration, Paris, 5/13
- Michael Jordan, Keynote Speaker, Amazon Machine Learning Conference, Seattle, 4/28/13
- Michael Jordan, Keynote Speaker, Bayesian Nonparametrics Workshop, Amsterdam, 6/10/13
- Michael Jordan, Invited Speaker, Workshop on High-Dimensional Statistics, Moscow, 6/26/13
- Michael Jordan, Distinguished Lecture, Department of Statistics, University of Oxford, 5/7/13
- Michael Jordan, Colloquium, Department of Statistics, University of Cambridge, 5/10/13
- Michael Jordan, Invited Speaker, GdR ISIS Conference, Telecom ParisTech, Paris 5/16/13
- Matthieu Solnon, "Analysis of the oracle risk in multi-task kernel ridge regression", Colloque Statistique Mathématique et Applications, Fréjus, France.
- Mark Schmidt, "Opening up the black box: Faster methods for non-smooth and big-data optimization problems". Invited talk at DeepMind Technologies, London (June 2013).
- Mark Schmidt, "Linearly-Convergent Stochastic-Gradient Methods". Invited talk at Paris 6, Paris (June 2013).
- Mark Schmidt, "Minimizing Finite Sums with the Stochastic Average Gradient Algorithm". "Invited" talk at ICCOPT, Lisbon (July 2013).
- Edouard Grave, Alpage, Inria / Paris 7, May 2013
- Edouard Grave, Criteo, September 2013
- Edouard Grave, Laboratoire de Science Cognitive et Psycholinguistique, EHESS / ENS / CNRS, November 2013
- Alexandre d'Aspremont, "Convex Relaxations for Permutation Problems"
  - Workshop on Succinct Data Representations and Applications, Simons Institute, Berkeley, Sept. 2013.
  - Workshop MAORI, Ecole Polytechnique, Nov. 2013.
- Alexandre d'Aspremont, "Phase Retrieval, MAXCUT and Complex Semidefinite Programming"
  - GdT CEREMADE, Paris Dauphine, April 2013.
  - Journée du GdR ISIS, Telecom, May 2013.
  - Journée du GdR MOA, June 2013.
- Alexandre d'Aspremont,"Approximation Bounds for Sparse PCA"
  - Workshop on Structured families of functions and applications, Oberwolfach, February 2013.
  - PACM seminar, Princeton, USA, February 2013.
  - Séminaire ENSAE, France, April 2013.
  - Big Data workshop, IHP, May 2013.
- Alexandre d'Aspremont, "An Optimal Affine Invariant Smooth Minimization Algorithm", International Workshop on Statistical Learning, Moscow, June 2013.
- F. Bach: Optimization and Statistical Learning, January 6 - 11, 2013. Les Houches, France (Invited presentation)
- F. Bach: international biomedical and astronomical signal processing (BASP) Frontiers workshop, January 2013 (Invited presentation)
- F. Bach: Convex Relaxation Methods for Geometric Problems in Scientific Computing, IPAM, Los Angeles, February 2013 (Invited presentation)
- F. Bach: Nonsmooth optimization in machine learning. March 04, 2013, University of Liège (Invited presentation)

- F. Bach: Microsoft Research Machine Learning Summit: April 22-24, 2013 (Invited presentation)
- F. Bach: International Workshop on Advances in Regularization, Optimization, Kernel Methods and Support Vector Machines: theory and applications, July 8 - 10, 2013, Leuven, Belgium (Invited presentation)
- F. Bach: European Conference on Dana Analysis, Luxembourg, July 2013 (Invited presentation)
- F. Bach: European Meeting of Statisticians (EMS), Budapest, Hungary, 20-25 July 2013 (Invited presentation)
- F. Bach: Fourth Cargese Workshop on Combinatorial Optimization. Institut d'Etudes Scientifiques de Cargèse, Corsica (France). September 30 - October 5, 2013 (Invited presentation)
- F. Bach: 9èmes Journées Nationales de la Recherche en Robotique, Annecy, October 16-18, 2013 (invited presentation)
- F. Bach: Radboud University, Nijmegen, Nederlands, November 29, 2013 (Seminar)
- F. Bach: GlobalSIP: IEEE Global Conference on Signal and Information Processing, December 3-5, 2013 (invited presentation)
- F. Bach: NIPS workshops, december 2013 (2 invited presentations)

9.2. Teaching - Supervision - Juries

9.2.1. Teaching

Licence: F. Bach, G. Obozinski, R. Lajugie: "Apprentissage statistique", 35h, Ecole Normale Supérieure, Filière "Math-Info", première année.

Mastère: S. Arlot and F. Bach, "Statistical learning", 24h, Mastère M2, Université Paris-Sud, France.

Mastère: F. Bach, G. Obozibski, Introduction aux modèles graphiques (30h), Master MVA (Ecole Normale Supérieure de Cachan).

Master: S. Arlot and F. Bach, "Statistical learning", 24h, Mastère M2, Université Paris-Sud, France.

Doctorat: S. Arlot, "Classification and statistical machine learning", 1h tutorial for the CEMRACS 2013, Marseille, France.

Licence : A. d'Aspremont, "Optimisation", 36h, L3, ENSAE, France.

Master M2: A. d'Aspremont, "Convex Optimization, Algorithms and Applications", 27h, M2, ENS Cachan, France.

Master M2: A. d'Aspremont, "Optimisation et simulation numérique.", 14h, M2, Paris Sud (Orsay), France.

9.2.2. Supervision

PhD: Matthieu Solnon, "Multi-task statistical learning", UPMC, November 25, 2013. Advisors: S. Arlot and F. Bach.

PhD in progress: Fajwel Fogel, "Optimisation et Apprentissage", September 2012, A. d'Aspremont.

PhD in progress: Bamdev Mishra

PhD in progress: Loic Landrieu

PhD in progress: Sesh Kumar, "Optimization and submodular functions", May 2012, F. Bach.

PhD in progress: Edouard Grave, "A Markovian approach to distributional semantics", F. Bach, G. Obozinski (defended January 20, 2014).

PhD in progress: Anil Nelakanti, "Structured sparsity and language models", F. Bach, G. Obozinski (to be defended February 11, 2014).

PhD in progress: Rémi Lajugie, September 2012, S. Arlot and F. Bach.

9.2.3. Juries

- A. d'Aspremont was a member of the PhD committee of Nicholas Boumal's thesis at the Université Catholique de Louvain.
- F. Bach was a member of the PhD committee of Clément Calauzènes (UPMC), Azadeh Khaleghi (Inria Lille), Yao-Lian Yu (University of Alberta).
- F. Bach was a member of the HDR committee of Ivan Laptev (Inria-ENS), Pawan Kumar (Ecole Centrale).

9.3. Popularization

- A. d'Aspremont, Journée Big Data, ENSAI, Rennes, November 2013.

10. Bibliography

Publications of the year

Doctoral Dissertations and Habilitation Theses

[1] E. GRAVE. , *A Markovian approach to distributional semantics*, Université Pierre et Marie Curie - Paris VI, January 2014, http://hal.inria.fr/tel-00940575

[2] M. SOLNON. , *Apprentissage statistique multi-tâches*, Université Pierre et Marie Curie - Paris VI, November 2013, http://hal.inria.fr/tel-00911498

Articles in International Peer-Reviewed Journals

[3] Z. HARCHAOUI, F. BACH, O. CAPPÉ, E. MOULINES. *Kernel-Based Methods for Hypothesis Testing: A Unified View*, in "IEEE Signal Processing Magazine", June 2013, vol. 30, n^{o} 4, pp. 87-97 [*DOI :* 10.1109/MSP.2013.2253631], http://hal.inria.fr/hal-00841978

[4] B. MISHRA, G. MEYER, F. BACH, R. SEPULCHRE. *Low-rank optimization with trace norm penalty*, in "SIAM Journal on Optimization", 2013, vol. 23, n^{o} 4, pp. 2124-2149 [*DOI :* 10.1137/110859646], http://hal.inria.fr/hal-00924110

[5] A. D'ASPREMONT, N. E. KAROUI. *Weak Recovery Conditions from Graph Partitioning Bounds and Order Statistics*, in "Mathematics of Operations Research", July 2013, vol. 38, n^{o} 2, Final version, http://pubsonline.informs.org/doi/abs/10.1287/moor.1120.0581, http://hal.inria.fr/hal-00907541

International Conferences with Proceedings

[6] A. AHMED, N. SHERVASHIDZE, S. NARAYANAMURTHY, V. JOSIFOVSKI, A. J. SMOLA. *Distributed Large-scale Natural Graph Factorization*, in "IW3C2 - International World Wide Web Conference", Rio de Janeiro, Brazil, May 2013, 37 p. , http://hal.inria.fr/hal-00918478

[7] F. BACH. *Sharp analysis of low-rank kernel matrix approximations*, in "International Conference on Learning Theory (COLT)", United States, 2013, http://hal.inria.fr/hal-00723365

[8] F. BACH, E. MOULINES. *Non-strongly-convex smooth stochastic approximation with convergence rate O(1/n)*, in "Neural Information Processing Systems (NIPS)", United States, 2013, http://hal.inria.fr/hal-00831977

[9] P. BOJANOWSKI, F. BACH, I. LAPTEV, J. PONCE, C. SCHMID, J. SIVIC. *Finding Actors and Actions in Movies*, in "ICCV 2013 - IEEE International Conference on Computer Vision", Sydney, Australia, IEEE, 2013, http://hal.inria.fr/hal-00904991

[10] M. CUTURI, A. D'ASPREMONT. *Mean Reversion with a Variance Threshold*, in "International Conference on Machine Learning", United States, October 2013, pp. 271-279, http://hal.inria.fr/hal-00939566

[11] M. EICKENBERG, F. PEDREGOSA, S. MEHDI, A. GRAMFORT, B. THIRION. *Second order scattering descriptors predict fMRI activity due to visual textures*, in "PRNI 2013 - 3nd International Workshop on Pattern Recognition in NeuroImaging", Philadelphia, United States, Conference Publishing Services, June 2013, http://hal.inria.fr/hal-00834928

[12] F. FOGEL, R. JENATTON, F. BACH, A. D'ASPREMONT. *Convex Relaxations for Permutation Problems*, in "Neural Information Processing Systems (NIPS) 2013", United States, August 2013, http://nips.cc/Conferences/2013/Program/speaker-info.php?ID=12863, http://hal.inria.fr/hal-00907528

[13] E. GRAVE, G. OBOZINSKI, F. BACH. *Hidden Markov tree models for semantic class induction*, in "CoNLL - Seventeenth Conference on Computational Natural Language Learning", Sofia, Bulgaria, 2013, http://hal.inria.fr/hal-00833288

[14] P. GRONAT, G. OBOZINSKI, J. SIVIC, T. PAJDLA. *Learning and calibrating per-location classifiers for visual place recognition*, in "CVPR 2013 - 26th IEEE Conference on Computer Vision and Pattern Recognition", Portland, United States, June 2013, http://hal.inria.fr/hal-00934332

[15] S. JEGELKA, F. BACH, S. SRA. *Reflection methods for user-friendly submodular optimization*, in "NIPS 2013 - Neural Information Processing Systems", Lake Tahoe, Nevada, United States, 2013, http://hal.inria.fr/hal-00905258

[16] S. LACOSTE-JULIEN, M. JAGGI, M. SCHMIDT, P. PLETSCHER. *Block-Coordinate Frank-Wolfe Optimization for Structural SVMs*, in "ICML 2013 International Conference on Machine Learning", Atlanta, United States, 2013, pp. 53-61, http://hal.inria.fr/hal-00720158

[17] S. LACOSTE-JULIEN, K. PALLA, A. DAVIES, G. KASNECI, T. GRAEPEL, Z. GHAHRAMANI. *SiGMa: Simple Greedy Matching for Aligning Large Knowledge Bases*, in "KDD 2013 - The 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining", Chicago, United States, August 2013, pp. 572-580 [*DOI :* 10.1145/2487575.2487592], http://hal.inria.fr/hal-00918671

[18] N. LE ROUX, F. BACH. *Local Component Analysis*, in "ICLR - International Conference on Learning Representations 2013", Scottsdale, United States, 2013, http://hal.inria.fr/inria-00617965

[19] A. NELAKANTI, C. ARCHAMBEAU, J. MAIRAL, F. BACH, G. BOUCHARD. *Structured Penalties for Log-linear Language Models*, in "EMNLP - Empirical Methods in Natural Language Processing - 2013", Seattle, United States, Association for Computational Linguistics, October 2013, pp. 233-243, http://hal.inria.fr/hal-00904820

[20] F. PEDREGOSA, M. EICKENBERG, B. THIRION, A. GRAMFORT. *HRF estimation improves sensitivity of fMRI encoding and decoding models*, in "3nd International Workshop on Pattern Recognition in NeuroImaging", Philadelphia, United States, May 2013, http://hal.inria.fr/hal-00821946

[21] E. RICHARD, F. BACH, J.-P. VERT. *Intersecting singularities for multi-structured estimation*, in "ICML 2013 - 30th International Conference on Machine Learning", Atlanta, United States, 2013, http://hal.inria.fr/hal-00918253

[22] G. RIGAILL, T. D. HOCKING, F. BACH, J.-P. VERT. *Learning Sparse Penalties for Change-Point Detection using Max Margin Interval Regression*, in "ICML 2013 - 30 th International Conference on Machine Learning", Atlanta, United States, Supported by the International Machine Learning Society (IMLS), May 2013, http://hal.inria.fr/hal-00824075

[23] T. SCHATZ, V. PEDDINTI, F. BACH, A. JANSEN, H. HERMANSKY, E. DUPOUX. *Evaluating speech features with the Minimal-Pair ABX task: Analysis of the classical MFC/PLP pipeline*, in "INTERSPEECH 2013 : 14th Annual Conference of the International Speech Communication Association", Lyon, France, 2013, pp. 1-5, http://hal.inria.fr/hal-00918599

[24] K. S. SESH KUMAR, F. BACH. *Convex Relaxations for Learning Bounded Treewidth Decomposable Graphs*, in "International Conference on Machine Learning", Atlanta, United States, 2013, Extended version of the ICML-2013 paper., http://hal.inria.fr/hal-00763921

Conferences without Proceedings

[25] E. GRAVE, G. OBOZINSKI, F. BACH. *Domain adaptation for sequence labeling using hidden Markov models*, in "New Directions in Transfer and Multi-Task: Learning Across Domains and Tasks (NIPS Workshop)", Lake Tahoe, United States, 2013, http://hal.inria.fr/hal-00918371

Scientific Books (or Scientific Book chapters)

[26] F. BACH. , *Learning with Submodular Functions: A Convex Optimization Perspective*, Foundations and Trends in Machine Learning, Now Publishers, 2013, 228 p. [*DOI :* 10.1561/2200000039], http://hal.inria.fr/hal-00645271

Other Publications

[27] F. BACH. , *Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression*, October 2013, http://hal.inria.fr/hal-00804431

[28] F. BACH. , *Convex relaxations of structured matrix factorizations*, September 2013, http://hal.inria.fr/hal-00861118

[29] F. FOGEL, I. WALDSPURGER, A. D'ASPREMONT. , *Phase retrieval for imaging problems*, 2013, http://hal.inria.fr/hal-00907529

[30] R. GRIBONVAL, R. JENATTON, F. BACH, M. KLEINSTEUBER, M. SEIBERT. , *Sample Complexity of Dictionary Learning and other Matrix Factorizations*, December 2013, submitted, http://hal.inria.fr/hal-00918142

[31] R. LAJUGIE, S. ARLOT, F. BACH. , *Large-Margin Metric Learning for Partitioning Problems*, March 2013, http://hal.inria.fr/hal-00796921

[32] M. SCHMIDT, N. LE ROUX, F. BACH. , *Minimizing Finite Sums with the Stochastic Average Gradient*, September 2013, http://hal.inria.fr/hal-00860051

[33] M. SCHMIDT, N. LE ROUX. , *Fast Convergence of Stochastic Gradient Descent under a Strong Growth Condition*, August 2013, http://hal.inria.fr/hal-00855113

[34] K. S. SESH KUMAR, F. BACH. , *Maximizing submodular functions using probabilistic graphical models*, September 2013, http://hal.inria.fr/hal-00860575

[35] M. SOLNON. , *Comparison bewteen multi-task and single-task oracle risks in kernel ridge regression*, 2013, Submitted to the Electronic Journal of Statistics, http://hal.inria.fr/hal-00846715

[36] I. WALDSPURGER, A. D'ASPREMONT, S. MALLAT. , *Phase Recovery, MaxCut and Complex Semidefinite Programming*, 2013, Submitted revision, http://hal.inria.fr/hal-00907535

[37] A. D'ASPREMONT, M. JAGGI. , *An Optimal Affine Invariant Smooth Minimization Algorithm*, 2013, http://hal.inria.fr/hal-00907547

References in notes

[38] F. BACH. *Learning with Submodular Functions: A Convex Optimization Perspective*, in "ArXiv e-prints", 2011

[39] F. BACH, M. JORDAN. *Thin junction trees*, in "Adv. NIPS", 2002

[40] F. BACH, M. JORDAN. *Learning spectral clustering*, in "Adv. NIPS", 2003

[41] A. BAR-HILLEL, T. HERTZ, N. SHENTAL, D. WEINSHALL. *Learning a mahalanobis metric from equivalence constraints*, in "Journal of Machine Learning Research", 2006, vol. 6, n^{o} 1, 937 p.

[42] C. BISHOP, ET AL.. , *Pattern recognition and machine learning*, springer New York, 2006

[43] D. BLATT, A. O. HERO, H. GAUCHMAN. *A convergent incremental gradient method with a constant step size*, in "SIOPT", 2007, vol. 18, n^{o} 1, pp. 29–51

[44] Y. BOYKOV, O. VEKSLER, R. ZABIH. *Fast approximate energy minimization via graph cuts*, in "IEEE Trans. PAMI", 2001, vol. 23, n^{o} 11, pp. 1222–1239

[45] L. BURGET, P. MATEJKA, P. SCHWARZ, O. GLEMBEK, J. CERNOCKY. *Analysis of Feature Extraction and Channel Compensation in a GMM Speaker Recognition System*, in "IEEE Transactions on Audio, Speech and Language Processing", September 2007, vol. 15, n^{o} 7, pp. 1979-1986

[46] M. A. CARLIN, S. THOMAS, A. JANSEN, H. HERMANSKY. *Rapid evaluation of speech representations for spoken term discovery*, in "Proceedings of Interspeech", 2011

[47] Y.-W. CHANG, M. COLLINS. *Exact Decoding of Phrase-based Translation Models through Lagrangian Relaxation*, in "Proceedings of the Conference on Empirical Methods for Natural Language Processing", 2011, pp. 26–37

[48] A. CHECHETKA, C. GUESTRIN. *Efficient Principled Learning of Thin Junction Trees*, in "Adv. NIPS", 2007

[49] J. CHEN, A. K. GUPTA. , *Parametric Statistical Change Point Analysis*, Birkhäuser, 2011

[50] S. CHEN, R. ROSENFELD. *A survey of smoothing techniques for ME models*, in "IEEE Transactions on Speech and Audio Processing", 2000, vol. 8, n^{o} 1, pp. 37–50

[51] Y. CHENG. *Mean shift, mode seeking, and clustering*, in "IEEE Trans. PAMI", 1995, vol. 17, n^{o} 8, pp. 790–799

[52] C. I. CHOW, C. N. LIU. *Approximating discrete probability distributions with dependence trees*, in "IEEE Trans. Inf. Theory", 1968, vol. 14

[53] F. DE LA TORRE, T. KANADE. *Discriminative cluster analysis*, in "Proc. ICML", 2006

[54] F. DESOBRY, M. DAVY, C. DONCARLI. *An online kernel change detection algorithm*, in "IEEE Trans. Sig. Proc.", 2005, vol. 53, n^{o} 8, pp. 2961–2974

[55] B. EFRON, C. N. MORRIS. *Stein's paradox in statistics*, in "Scientific American", 1977, vol. 236, pp. 119–127

[56] T. EVGENIOU, C. A. MICCHELLI, M. PONTIL. *Learning Multiple Tasks with Kernel Methods*, in "Journal of Machine Learning Research", 2005, vol. 6, pp. 615–637

[57] P. FOUSEK, P. SVOJANOVSKY, F. GREZL, H. HERMANSKY. *New Nonsense Syllables Database – Analyses and Preliminary ASR Experiments*, in "Proceedings of the International Conference on Spoken Language Processing (ICSLP)", 2004, pp. 2004-29

[58] S. FUJISHIGE. , *Submodular Functions and Optimization*, Annals of Discrete Mathematics, Elsevier, 2005

[59] V. GOGATE, W. WEBB, P. DOMINGOS. *Learning Efficient Markov Networks*, in "Adv. NIPS", 2010

[60] J. GOODMAN. *A bit of progress in language modelling*, in "Computer Speech and Language", October 2001, pp. 403–434

[61] J. C. GOWER, G. J. S. ROSS. *Minimum spanning trees and single linkage cluster analysis*, in "Applied statistics", 1969, pp. 54–64

[62] T. D. HOCKING, G. SCHLEIERMACHER, I. JANOUEIX-LEROSEY, O. DELATTRE, F. BACH, J.-P. VERT. *Learning smoothing models of copy number profiles using breakpoint annotations*, in "HAL, archives ouvertes", 2012

[63] L. JACOB, F. BACH, J.-P. VERT. *Clustered Multi-Task Learning: A Convex Formulation*, in "Computing Research Repository", 2008, pp. -1–1

[64] W. JAMES, C. STEIN. *Estimation with quadratic loss*, in "Proceedings of the fourth Berkeley symposium on mathematical statistics and probability", 1961, vol. 1, n^{o} 1961, pp. 361–379

[65] R. JENATTON, J. MAIRAL, G. OBOZINSKI, F. BACH. *Proximal Methods for Hierarchical Sparse Coding*, in "Journal of Machine Learning Research", 2011, pp. 2297-2334

[66] R. KNESER, H. NEY. *Improved backing-off for m-gram language modeling*, in "Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing", 1995, vol. 1

[67] D. KOLLER, N. FRIEDMAN. , *Probabilistic graphical models: principles and techniques*, MIT press, 2009

[68] V. KOLMOGOROV, T. SCHOENEMANN. *Generalized sequential tree-reweighted message passing*, in "ArXiv e-prints", May 2012

[69] A. KRAUSE, C. GUESTRIN. *Submodularity and its Applications in Optimized Information Gathering*, in "ACM Transactions on Intelligent Systems and Technology", 2011, vol. 2, n^{o} 4

[70] H. LIN, J. BILMES. *A Class of Submodular Functions for Document Summarization*, in "Proc. NAACL/HLT", 2011

[71] D. LUENBERGER, Y. YE. , *Linear and nonlinear programming*, Springer Verlag, 2008

[72] N. A. MACMILLAN, C. D. CREELMAN. , *Detection theory: A user's guide*, Lawrence Erlbaum, 2004

[73] F. MALVESTUTO. *Approximating discrete probability distributions with decomposable models*, in "IEEE Trans. Systems, Man, Cybernetics", 1991, vol. 21, n^{o} 5

[74] A. F. T. MARTINS, N. A. SMITH, A. M. Q. PEDRO, M. A. T. FIGUEIREDO. *Structured sparsity in structured prediction*, in "Proceedings of the Conference on Empirical Methods for Natural Language Processing", 2011, pp. 1500–1511

[75] M. NARASIMHAN, J. BILMES. *PAC-learning bounded tree-width graphical models*, in "Proc. UAI", 2004

[76] G. NEMHAUSER, L. WOLSEY, M. FISHER. *An analysis of approximations for maximizing submodular set functions–I*, in "Mathematical Programming", 1978, vol. 14, n^{o} 1, pp. 265–294

[77] A. NEMIROVSKI, A. JUDITSKY, G. LAN, A. SHAPIRO. *Robust stochastic approximation approach to stochastic programming*, in "SIOPT", 2009, vol. 19, n^{o} 4, pp. 1574–1609

[78] A. NEMIROVSKI. *Efficient methods in convex programming*, in "Lecture notes", 1994

[79] Y. NESTEROV. , *Introductory lectures on convex optimization: A basic course*, Springer, 2004

[80] A. Y. NG, M. JORDAN, Y. WEISS. *On spectral clustering: Analysis and an algorithm*, in "Adv. NIPS", 2002

[81] B. ROARK, M. SARACLAR, M. COLLINS, M. JOHNSON. *Discriminative language modeling with conditional random fields and the perceptron algorithm*, in "Proceedings of the Association for Computation Linguistics", 2004

[82] L. SAUL, M. JORDAN. *Exploiting Tractable Substructures in Intractable Networks*, in "Adv. NIPS", 1995

[83] H. D. SHERALI, W. P. ADAMS. *A Hierarchy of Relaxations Between the Continuous and Convex Hull Representations for Zero-One Programming Problems*, in "SIAM J. Discrete Math.", 1990

[84] J. SHI, J. MALIK. *Normalized Cuts and Image Segmentation*, in "IEEE Trans. PAMI", 1997, vol. 22, pp. 888–905

[85] GSVS. SIVARAM, H. HERMANSKY. *Sparse Multilayer Perceptron for Phoneme Recognition*, in "IEEE Transactions on Audio, Speech, and Language Processing", 2012, vol. 20, n^{o} 1, pp. 23-29

[86] M. SOLNON, S. ARLOT, F. BACH. *Multi-task Regression using Minimal Penalties*, in "Journal of Machine Learning Research", September 2012, vol. 13, pp. 2773-2812

[87] M. SOLODOV. *Incremental gradient algorithms with stepsizes bounded away from zero*, in "Computational Optimization and Applications", 1998, vol. 11, n^{o} 1, pp. 23–35

[88] C. STEIN. *Inadmissibility of the usual estimator for the mean of a multivariate normal distribution*, in "Proceedings of the Third Berkeley symposium on mathematical statistics and probability", 1956, vol. 1, n^{o} 399, pp. 197–206

[89] T. SZÁNTAI, E. KOVÁCS. *Discovering a junction tree behind a Markov network by a greedy algorithm*, in "ArXiv e-prints", April 2011

[90] P. TSENG. *An incremental gradient(-projection) method with momentum term and adaptive stepsize rule*, in "SIOPT", 1998, vol. 8, n^{o} 2, pp. 506-531

[91] I. TSOCHANTARIDIS, T. HOFMANN, T. JOACHIMS, Y. ALTUN. *Support Vector Machine Learning for Interdependent and Structured Output Spaces*, in "Proc. ICML", 2004

[92] S. VARGAS, P. CASTELLS, D. VALLET. *Explicit relevance models in intent-oriented information retrieval diversification*, in "Proceedings of the 35th ACM SIGIR International Conference on Research and development in information retrieval", Portland, Oregon, USA, SIGIR'12, ACM, 2012, pp. 75-84, http://doi.acm.org/10.1145/2348283.2348297

[93] M. WAINWRIGHT, M. JORDAN. *Graphical models, exponential families, and variational inference*, in "Found. and Trends in Mach. Learn.", 2008, vol. 1, n^{o} 1-2

[94] F. WOOD, C. ARCHAMBEAU, J. GASTHAUS, J. LANCELOT, Y.-W. TEH. *A Stochastic Memoizer for Sequence Data*, in "Proceedings of the 26th International Conference on Machine Learning", 2009

[95] E. P. XING, A. Y. NG, M. JORDAN, S. RUSSELL. *Distance metric learning with applications to clustering with side-information*, in "Adv. NIPS", 2002

[96] P. ZHAO, G. ROCHA, B. YU. *The composite absolute penalties family for grouped and hierarchical variable selection*, in "The Annals of Statistics", 2009, vol. 37(6A), pp. 3468-3497