Machine learning

Centre International de Mathématiques et d'Informatique
TOULOUSE

# Workshop 1:

# Optimization in machine learning, vision and image processing

Université Paul Sabatier, Toulouse,
October 6th-7th 2015

# About the CIMI Machine Learning thematic trimester

From September to December 2015, the International Centre for Mathematics and Computer Science (CIMI) organizes a thematic trimester on Machine Learning.

The goal of this trimester is to propose a series of scientific and pedagogic events reflecting common interests of the two laboratories that founded CIMI:Institut de Mathématiques de Toulouse (IMT) and Institut de Recherche en Informatique de Toulouse (IRIT).

The focus will be set in particular on three axes:

- optimization problems in machine learning,
- learning with structured data and natural language,
- sequential learning.

The trimester will start with a summer school and will continue with three thematic workshops.

- Summer school - 14th to 18th September 2015
- Workshop 1 - Optimization in machine learning, vision and image processing - 6th to 7th October 2015
- Workshop 2 - Sequential Learning and Applications - 9th to 10th November 2015
- Workshop 3 - Learning with Structured Data and Natural Language - 9th to 11th December 2015

## Scientific committee

- Francis Bach, Ecole Normale Supérieure
- Sébastien Bubeck, Microsoft Research
- Nicolo Cesa-Bianchi, Universitá degli Studi di Milano
- Rémi Gribonval, lRISA Rennes
- Marc Sebban, University Jean Monnet
- Noah Smith, Carnegie Mellon University
- Johan Suykens, KU Leuven
- Marc Teboulle, Tel-Aviv University



Machine learning

cimi
Centre International de Mathématiques et d'Informatique
TOULOUSE

# CIMI Semester on Machine Learning

## **Workshop 1:** Optimization in machine learning, vision and image processing

October 6th-7th 2015

The scientific program includes invited talks describing the most recent advances on the theoretical properties, the numerical resolution and the modelling of optimization problems to solve problems in machine learning, vision and image processing. The typical problems optimize large vectors, matrices or tensors according to criteria involving L2, L1 and other sparsity enforcing terms, nuclear norms, Wasserstein distance and/or non-convex criteria and/or solving bi-level problems.

## List of Speakers

- Amir Beck, Technion
- Marco Cuturi, Kyoto University
- Martin Jaggi, ETH Zürich
- Joseph Landsberg, Texas A&M University
- Lek Heng Lim, University of Chicago
- Julien Mairal, INRIA Grenoble
- Eric Moulines, Telecom Paristech
- Gabriel Peyré, University of Paris-Dauphine
- Jean Ponce, Ecole Normale Supérieure
- Saverio Salzo, University of Genoa
- Johan Suykens, KU Leuven
- Marc Teboulle, Tel Aviv University
- Yuning Yang, KU Leuven




# Schedule

| | | |
|---|---|---|
| **Tuesday, October 6th** | | |
| | | **Chairman: F. Malgouyres** |
| 8:45 | | Welcome |
| 9:15 | J. Suykens | Learning with primal and dual model representations |
| 10:00 | J. Mairal | A Universal Catalyst for First-Order Optimization |
| 10:45 | | Coffee Break |
| 11:00 | J. Landsberg | On the complexity of matrix multiplication and other tensors |
| 11:45 | | Lunch |
| | | **Chairman: J. Suykens** |
| 13:30 | Lek Heng Lim | Higher order cone programming |
| 14:15 | Yuning Yang | Rank-One Tensor Updating Algorithms for A Class of Low Rank Tensor Learning (Optimization) Problems |
| 15:00 | | Coffee Break |
| | | **Chairman: J. Mairal** |
| 15:30 | E. Moulines | How to sample efficiently over high-dimensional space ? |
| 16:15 | M. Jaggi | L1-Regularized Distributed Optimization: A Communication-Efficient Primal-Dual Framework |
| 17:00 | | End |
| **Wednesday, October 7th** | | |
| | | **Chairman: S. Gadat** |
| 9:15 | J. Ponce | Weakly Supervised Image and Video Interpretation |
| 10:00 | | Coffee Break |
| 10:15 | M. Cuturi | New Approaches to Learn with Probability Measures using Fast Optimal Transport |
| 11:00 | G. Peyré | Entropic Approximation of Wasserstein Gradient Flow |
| 11:45 | | Lunch |
| | | **Chairman: J.-B. Hiriart Urruty** |
| 13:30 | M. Teboulle | Simple Minimization Algorithms for Difficult Optimization Problems Illustrated |
| 14:15 | | Coffee Break |
| 14:45 | S. Salzo | Consistent Learning by Composite Proximal Thresholding |
| 15:30 | A. Beck | Primal and Dual Variables Decomposition Methods in Convex Optimization |
| 16:15 | | End |

Machine learning

Centre International de Mathématiques et d'Informatique
TOULOUSE

# List of abstracts

**Amir Beck** *Primal and Dual Variables Decomposition Methods in Convex Optimization*

We consider the rates of convergence of several decomposition methods, which are based on either exact block minimization or on the block proximal gradient operator. We will then discuss the model of minimizing a strongly convex function and a term comprising the sum of extended real-valued convex functions (atoms). We derive several dual-based variables decomposition methods that take into account only one of the atoms and analyze their rate of convergence through a simple primal-dual formula. Finally, the effectiveness of the methods are illustrated through some image denoising problems.

**Marco Cuturi** *New Approaches to Learn with Probability Measures using Fast Optimal Transport* Optimal transport distances (a.k.a Wasserstein distances or Earth Mover's distances, EMD) define a geometry for empirical measures supported on a metric space. After reviewing the basics of the optimal transport problem, I will show how an adequate regularization of that problem can result in substantially faster computations [a]. I will then show how this regularization can enable several applications of optimal transport to learn from probability measures, from the computation of barycenters [b,c,d,e] to that of dictionaries [f] or PCA [g], all carried out using the Wasserstein geometry.

Papers:

[a] MC, Sinkhorn Distances: Lightspeed Computation of Optimal Transport, NIPS 2013.

[b] MC, A. Doucet, Fast Computation of Wasserstein Barycenters, ICML 2014.

[c] J.D. Benamou, G. Carlier, MC, L. Nenna, G. Peyré, Iterative Bregman Projections for Regularized Transportation Problems, to appear in SIAM Journ. on Scientific Computing.

[d] J. Solomon, F. de Goes, G. Peyré, MC, A. Butscher, A. Nguyen, T. Du, L. Guibas. Convolutional Wasserstein Distances: Efficient Optimal Transportation on Geometric Domains, SIGGRAPH 2015.

[e] A. Gramfort, G. Peyré , MC, Fast Optimal Transport Averaging of Neuroimaging Data, IPMI 2015.

[f] MC, G. Peyré, A. Rolet, A Smoothed Dual Formulation for Variational Wasserstein Problems, arxiv:1503.02533, 2015.

[g] V. Séguy, MC An Algorithmic Approach to Compute Principal Geodesics in the Wasserstein Space, arXiv:1506.07944, 2015

**Martin Jaggi** *L1-Regularized Distributed Optimization: A Communication-Efficient Primal-Dual Framework*

Despite the importance of sparsity in many big data applications, there are few existing methods for efficient distributed optimization of sparsely-regularized objectives. In this paper, we present a communication-efficient framework for L1-regularized optimization in the distributed environment. By taking a non-traditional view of classical objectives as part of a more general primal-dual setting - and swapping the roles of regularizers and losses - we obtain a new class of methods that can be efficiently distributed and are applicable to many common L1-regularized regression and classification objectives, such as Lasso, sparse logistic regression, and elastic net regularized problems. We provide convergence guarantees for this framework and demonstrate strong empirical performance as compared to other state-of-the-art methods on several real-world distributed datasets.


**Joseph Landsberg** *On the complexity of matrix multiplication and other tensors*

I will discuss recent developments in obtaining lower bounds for rank and border rank of tensors with an emphasis on applications to the matrix multiplication tensor.

**Lek Heng Lim** *Higher order cone programming*

We introduce a family of cone programming problems that interpolates between LP, SOCP, and SDP. We show that there is a natural family of kth order cones that may be realized either as cones of n-by-n symmetric matrices or as cones of n-variate even degree polynomials. The cases k = 1, 2, n, correspond precisely to the nonnegative orthant, the Lorentz cone, and the semidefinite cone respectively. Linear optimization over these cones then correspond to LP, SOCP, SDP; alternatively, in the language of polynomial optimization, they correspond to DSOS, SDSOS, SOS. For general values of k between 3 and n - 1, we obtain new cone programming problems that we call kOCP's. We will discuss the properties of kOCP and see how the assembly of LP, SOCP, and SDP in such a hierarchy shed light on their relative computational efficiencies. This is joint work with Lijun Ding (Chicago) and is inspired by recent work of Anirudha, Ahmadi, and Tedrake.

**Julien Mairal** *A Universal Catalyst for First-Order Optimization*

We introduce a generic scheme for accelerating first-order optimization methods in the sense of Nesterov. Our approach consists of minimizing a convex objective by approximately solving a sequence of well-chosen auxiliary problems, leading to faster convergence. This strategy applies to a large class of algorithms, including gradient descent, block coordinate descent, SAG, SAGA, SDCA, SVRG, Finito/MISO, and their proximal variants. For all of these approaches, we provide acceleration and explicit support for non-strongly convex objectives. In addition to theoretical speed-up, we also show that acceleration is useful in practice, especially for ill-conditioned problems where we measure dramatic improvements. This is a joint work with Hongzhou Lin and Zaid Harchaoui.

**Eric Moulines** *How to sample efficiently over high-dimensional space ?*

Sampling over high-dimensional space has become a prerequisite in the applications of Bayesian statistics to machine learning problem. In many situations of interest, the log-posterior distribution is concave. The likelihood part is generally smooth and gradient Lipshitz while the prior is concave but typically not smooth (the archetypical problem is the LASSO or the elastic-net penalty, but many other problems can be cast into this framework). We will describe methods to sample such distributions, which are adapted from the state-of-the-art optimization procedures which have been developed in this context. We will also provide convergence in Wasserstein distance to the equilibrium, showing explicitly the dependence in the dimension of the parameter space and the sparsity (effective dimension of the model).

Machine learning

Centre International de Mathématiques et d'Informatique
TOULOUSE

## Gabriel Peyré *Entropic Approximation of Wasserstein Gradient Flow*

In this talk I will detail a novel numerical scheme to approximate gradient flows for optimal transport (i.e. Wasserstein) metrics. These flows have proved useful to tackle theoretically and numerically non-linear diffusion equations that model for instance porous media or crowd evolutions. A bottleneck of these approaches is the high computational load induced by the resolution of each step. Indeed, this corresponds to the resolution of a convex optimization problem involving a Wasserstein distance to the previous iterate. Following several recent works on the approximation of Wasserstein distances, I consider a discrete flow induced by an entropic regularization of the transportation coupling. This entropic regularization allows one to trade the initial Wasserstein fidelity term for a Kullback-Leibler divergence, which is easier to deal with numerically. We show how Kullback-Leibler first order proximal schemes, and in particular DykstraâĂŹs algorithm, can be used to compute each step of the regularized flow. The resulting algorithm is both fast, parallelizable and versatile, because it only requires multiplications by the Gibbs kernel exp(-c/gamma) where c is the ground cost and gamma>0 the regularization strength. On Euclidean domains discretized on an uniform grid, this corresponds to a linear filtering (for instance a Gaussian filtering when c is the squared Euclidean distance) which can be computed in nearly linear time. On more general domains, such as (possibly non-convex) shapes or on manifolds discretized by a triangular mesh, following a recently proposed numerical scheme for optimal transport, this Gibbs kernel multiplication is approximated by a short-time heat diffusion. We show numerical illustrations of this method to approximate crowd motions on complicated domains as well as non-linear diffusions with spatially-varying coefficients.

## Jean Ponce *Weakly Supervised Image and Video Interpretation*

This talk addresses the problem of understanding the visual content of images and videos using a weak form of supervision, such as the fact that multiple images contain instances of the same objects, or the textual information available in television or film scripts. I will discuss several instances of this problem, including multi-class image cosegmentation, the joint localization and identification of movie characters and their actions, and the assignment of action labels to video frames using temporal ordering constraints. All these problems can be tackled using a discriminative clustering framework, and I will present the underlying models, appropriate relaxations of the corresponding combinatorial optimization problems associated with learning these models, and efficient algorithms for solving the corresponding convex optimization problems. I will also present experimental results on standard image benchmarks and feature-length films.

Joint work with Piotr Bojanowski, Armand Joulin, Remi Lajugie, Francis Bach, Minsu Cho, Suha Kwak, Ivan Laptev, Cordelia Schmid, and Josef Sivic.

## Saverio Salzo *Consistent Learning by Composite Proximal Thresholding*

I will consider random design least-squares regression with prediction functions which are linear combination of elements of a possibly infinite-dimensional dictionary. I will propose a flexible composite regularization model, which makes it possible to apply various priors to the coefficients of the prediction function, including hard constraints. I will first address the consistency of the estimators obtained by minimizing the regularized empirical risk. Then, I will present an error-tolerant composite proximal thresholding algorithm for computing such estimators.


**Johan Suykens** *Learning with primal and dual model representations*

While in parametric models sparsity is achieved by regularization, in kernel-based models it is obtained by the choice of an appropriate loss function. We propose a general framework for supervised and unsupervised learning, and beyond, which enables to conceive these different paradigms within a unified setting of optimization modelling. An example is to use kernel spectral clustering as a core model with feature map and kernel-based representations, for hierarchical image segmentation. These representations relate to primal and Lagrange dual problems, respectively. It is shown how this enables to incorporate prior knowledge, obtain sparse representations by approximate feature maps and develop large scale algorithms.

**Marc Teboulle** *Simple Minimization Algorithms for Difficult Optimization Problems Illustrated*

Most scientific and engineering problems are challenged by the fact they involve functions of a very large number of variables. In addition to the numerical difficulties due to the so-called curse of dimensionality, the resulting optimization models are often non-smooth and even non-convex. In this talk we review some fundamental approaches and recent results, illustrating how the presence of such difficulties may be handled. A particular highlight is the central role played by problems data information and structures which can be beneficially exploited both in theory and practice.

**Yuning Yang** *Rank-One Tensor Updating Algorithms for A Class of Low Rank Tensor Learning (Optimization) Problems*

Low rank tensor learning, such as low rank tensor approximation, tensor completion, as well as multilinear multitask learning, draws much attention in recent years. In this talk, we introduce algorithms, based on rank-one tensor updating, for solving the problems. At each iteration of the proposed algorithms, the main cost is only to compute a rank-one tensor, by approximately solving a tensor spectral norm problem. Comparing with matrix SVD based state-of-the-art methods, computing the rank-one tensor can be much cheaper, resulting into the efficiency of the proposed methods. Linear convergence rate is then established, either with a convex or nonconvex cost function. Experiments on (robust) tensor completion and multilinear multitask learning demonstrate the efficiency and effectiveness of the proposed algorithms.


## About CIMI

CIMI stands for Centre International de Mathématiques et Informatique de Toulouse and it is one of the Excellence projects selected by the ANR for the period 2012-2020.

CIMI brings together the teams of the Institut de Mathématiques de Toulouse (IMT) and the Institut de Recherche en Informatique de Toulouse (IRIT).

The CIMI aims at becoming an international reference in mathematics, computer science and their interactions.

The program will attract high-level scientists and students from around the world. It includes actions towards attractiveness, such as Excellence Chairs for long-term visitors, grants for Doctoral and Post-Doctoral students, as well as fellowships for Master students. Attractiveness is further enhanced with thematic trimesters organized within CIMI on specific topics including courses, seminars and workshops.

The innovative tools developed at CIMI will also have a strong economic impact on the region and will profit its industrial partners in major technological areas, making CIMI a strategic partner of the social and economic world.