ABSTRACTS

Pierre Alquier

On the Properties of Variational Approximations of Gibbs Posteriors.

PAC-Bayesian bounds are useful tools to control the prediction risk of aggregated estimators. When dealing with the exponentially weighted aggregate (EWA), these bounds lead in some settings to the proof that the predictions are minimax-optimal. EWA is usually computed through Monte Carlo methods. However, in many practical applications, the computational cost of Monte Carlo methods is prohibitive. It is thus tempting to replace these by (faster) optimization algorithms that aim at approximating EWA: we will refer to these methods as variational Bayes (VB) methods.

In this talk I will show, thanks to a PAC-Bayesian theorem, that VB approximations are well founded, in the sense that the loss incurred in terms of prevision risk is negligible in some classical settings such as linear classification, ranking... These approximations are implemented in the R package pac-vb (written by James Ridgway) that I will briefly introduce. I will especially insist on the the proof of the PAC-Bayesian theorem in order to explain how this result can be extended to other settings.

Joint work with James Ridgway (Bristol) and Nicolas Chopin (ENSAE).

Alexandre d'Aspremont

Renegar's Condition Number and Compressed Sensing Performance.

Renegar's condition number is a data-driven computational complexity measure for convex programs, generalizing classical condition numbers in linear systems. We provide evidence that for a broad class of compressed sensing problems, the worst case value of this algorithmic complexity measure taken over all signals matches the restricted eigenvalue of the observation matrix, which controls compressed sensing performance. This means that, in these problems, a single parameter directly controls computational complexity and recovery performance.

Joint work with Vincent Roulet and Nicolas Boumal.
Preprint: http://arxiv.org/abs/1506.03295

Quentin Berthet

Trade-offs in Statistical Learning

I will explore the notion of constraints on learning procedures, and discuss the impact that they can have on statistical precision. This is inspired by real-life concerns such as limits on time for computation, on reliability of observations, or communication between agents. I will show how these constraints can be shown to have a concrete cost on the statistical performance of these procedures, by describing several examples.

Alain Celisse

Using kernels to detect abrupt changes in time series

In this talk we discuss the change-point detection problem when dealing with complex data. Our goal is to present a new procedure involving positive semidefinite kernels and allowing to detect abrupt changes arising in the full distribution of the observations along the time (and not only in their means). This two-stage procedure is based first on dynamic programming, and second on a new l_0-type penalty derived from a non-asympsotic model selection result applying to vectors in a reproducing kernel Hilbert space. Since our procedure relies on the dynamic programming algorithm, which induces a high computational complexity at the first step, we will also discuss an improved version of this first step allowing to achieve a complexity of $O(n^2)$ in time and $O(n)$ in space. Finally, we will illustrate the behavior of our kernel change-point procedure on a wide range of simulated data. In particular we empirically validate our penalty since the resulting penalized criterion recovers the true (number of) change-points with high probability. We also infer the influence of the kernel on the final results in practice.

Rémi Gribonval

Projections, Learning, and Sparsity for Efficient Data Processing

The talk will discuss recent generalizations of sparse recovery guarantees and compressive sensing to the context of machine learning. Assuming some "low-dimensional model" on the probability distribution of the data, we will see that in certain scenarios it is indeed (empirically) possible to compress a large data-collection into a reduced representation, of size driven by the complexity of the learning task, while preserving the essential information necessary to process it. Two case studies will be given: compressive clustering, and compressive Gaussian Mixture Model estimation, with an illustration on large-scale model-based speaker verification. Time allowing, some recent results on compressive spectral clustering will also be discussed.

Emilie Kaufmann

Optimal Best Arm Identification with Fixed Confidence

This talk proposes a complete characterization of the complexity of best-arm identification in one-parameter bandit models. We first give a new, tight lower bound on the sample complexity, that is the total number of draws of the arms needed in order to identify the arm with highest mean with a prescribed accuracy. This lower bound does not take an explicit form, but reveals the existence of a vector of optimal proportions of draws of the arms, that can be computed efficiently. We then propose a 'Track-and-Stop' strategy, whose sample complexity is proved to asymptotically match the lower bound. It consists in a new sampling rule, which tracks the optimal proportions of arm draws, and a stopping rule for which we propose several interpretations and that can be traced back to Chernoff (1959).

Vianney Perchet

Highly-Smooth Zero-th Order Online Optimization

We consider online convex optimization with noisy zero-th order information, that is noisy function evaluations at any desired point. We focus on problems with high degrees of smoothness, such as online logistic regression. We show that as opposed to gradient-based algorithms, high-order smoothness may be used to improve estimation rates, with a precise dependence on the degree of smoothness and the dimension. In particular, we show that for infinitely differentiable functions, we recover the same dependence on sample size as gradient-based algorithms, with an extra dimension-dependent factor. This is done for convex and strongly-convex functions in constrained or global optimization (with either one point or two points noisy evaluations of the functions). Joint work with F. Bach.

Garvesh Raskutti

Algorithmic and statistical perspectives of randomized sketching for ordinary least-squares

In large-scale data settings, randomized 'sketching' has become an increasingly popular tool. In the numerical linear algebra literature, randomized sketching based on either random projections or sub-sampling has been shown to achieve optimal worst-case error. In particular the sketched ordinary least-squares (OLS) solution and the CUR decomposition have been shown to achieve optimal approximation error bounds in a worst-case setting. However, until recently there has been limited work on consider the performance of the OLS estimator under a statistical model using statistical metrics. In this talk I present some recent results which address both the performance of sketching in the statistical setting, where we assume an underlying statistical model and show that many of the existing intuitions and results are quite different from the worst-case algorithmic setting.

Ohad Shamir

Trade-offs in Distributed Learning

In many large-scale applications, learning must be done on training data which is distributed across multiple machines. This presents an important challenge, with multiple trade-offs between optimization accuracy, statistical performance, communication cost, and computational complexity. In this talk I'll describe some recent and upcoming results about distributed convex learning and optimization, including algorithms as well as fundamental performance barriers.

Silvia Villa

Generalization properties of multiple passes stochastic gradient method

The stochastic gradient method has become an algorithm of choice in machine learning, because of its simplicity and small computational cost, especially when dealing with big data sets. Despite its widespread use, the generalization properties of the variants of stochastic gradient method used in practice are relatively little understood. Most previous works consider generalization properties of SGM with only one pass over the data, while in practice

multiple passes are usually considered. The effect of multiple passes has been studied extensively for the optimization of an empirical objective, but the role for generalization is less clear. In this talk, we start filling this gap studying the generalization properties of multiple passes stochastic gradient method for least square regression in an abstract non parametric setting. We show that, if all other parameters are fixed a priori, the number of passes over the data indeed acts as a regularization parameter. The obtained bounds are sharp and matches those obtained with other regularized techniques such as ridge regression.