International Workshop on Statistical Learning

Program & Abstracts

June 26-28, Moscow

Wednesday, June 26/2013

9:50 - 10:00 Opening

10:00 - 10:30 Michael Jordan (University of California, Berkeley, USA) - Matrix Concentration Inequalities via the Method of Exchangeable Pairs

We develop a theoretical framework for establishing concentration inequalities for non commutative operators, focusing specifically for the spectral norm of random matrices. Our work reposes on Stein's method of exchangeable pairs, as elaborated by Chatterjee, and it provides a very different approach to concentration than that provided by the classical large deviation argument, which relies strongly on commutativity. When applied to a sum of independent random matrices, our approach yields matrix general izations of the classical inequalities due to Hoeffding, Bernstein, Khintchine, and Rosenthal. The same technique delivers bounds for sums of dependent random matrices and more general matrix valued functions of dependent random variables. [Joint work with Lester Mackey, Richard Chen, Brendan Farrell, and Joel Tropp]

10:30 - 11:00 Ankur Moitra (Institute of Advanced Study, USA) - Disentangling Mixtures of Gaussians

Given data drawn from a mixture of multivariate Gaussians, a basic problem is to accurately estimate the mixture parameters. We provide a polynomial-time algorithm for this problem for any fixed number k of Gaussians in n dimensions (even if they overlap), with provably minimal assumptions on the Gaussians and polynomial data requirements.
In statistical terms, our estimator converges at an inverse polynomial rate, and no such estima-

tor (even exponential time) was known for this problem (even in one dimension, restricted to two Gaussians). Our algorithm reduces the n-dimensional problem to the one dimensional problem, where the method of moments is applied. Additionally, in order to prove correctness for our univariate learning algorithm, we develop a novel explanation for why the method of moments (due to Pearson in 1894) works based on connections to algebraic geometry. [Joint work with Adam Tauman Kalai and Gregory Valiant. See also independent work of Mikhail Belkin and Kaushik Sinha]

11:00 - 11:30 Coffee break

11:30 - 12:00 Stephane Mallat (Ecole Normale Supérieure, Paris, France) - Deep Neural Network Learning by Scattering

Deep neural networks are remarkably successful hybrid classifiers, first trained on large data bases of unlabeled examples, and then optimized with a discriminative supervised classifier. They provide state of the art results in computer vision, speech recognition, music and bio-medical classification, with little mathematical understanding of their performance. We introduce a mathematical model of deep neural networks with scattering transforms, which cascade complex valued unitary operators and a contractive modulus. In this framework, unsupervised learning amounts to optimize a contraction of the space, while maximizing the volume occupied by representations of unlabeled examples. These deep scattering provides new models of stochastic processes, whose properties are analyzed. Wavelet unitary operators appear to be nearly optimal for the first network layers of many audio and image classifiers. Applications will be discussed and shown on images and sounds.

12:00 - 12:30 Alexandre D'Aspremont (CMAP Polytechnique, France) - An Optimal Affine Invariant Smooth Minimization Algorithm

We formulate an affine invariant implementation of the algorithm in Nesterov (1983). We show that the complexity bound is then proportional to a affine invariant regularity constants defined with respect to the Minkowski gauge of the feasible set. We then discuss implications in the design and efficient implementation of accelerated first-order methods. [Joint work with Martin Jaggi]

12:30 - 14:00 Lunch

14:00 - 14:30 Elmar Diederichs (PreMoLab MIPT, Russia) - The Development of Semidefinite Sparse Component Analysis

Sparse non-Gaussian component analysis is an unsupervised linear method of extracting any structure from high-dimensional distributed data based on estimating a low-dimensional non-Gaussian data component. In this paper we discuss a new approach with known a priori reduced

dimension to direct estimation of the projector on the target space using semidefinite programming. The new approach avoids the estimation of the data covariance matrix and overcomes the traditional separation of element estimation of the target space and target space reconstruction. This allows to reduced the sampling size while improving the sensitivity to a broad variety of deviations from normality. Moreover the complexity of the new approach is limited to $O(d \log d)$.

14:30 - 15:00 Alexandre Tsybakov (ENSAE-CREST, France) - Empirical Entropy, Minimax Regret and Minimax Risk

The study of the minimax risk over a class of functions and of a minimax regret based on the excess risk represents two parallel developments; the former has been mostly analyzed within Nonparametric Statistics, while the second – within Statistical Learning Theory. This talk aims to bring out a connection between these two objects. Considering the random design regression with square loss, we propose a method that aggregates empirical minimizers over appropriately chosen random subsets and we establish sharp oracle inequalities for its risk. We show that, under the ϵ^{-p} *growth of the empirical* ϵ*-entropy, the excess risk of the proposed method attains the rate* $n^{-\frac{2}{2+p}}$ *for* $p \in (0, 2]$ *and* $n^{-1/p}$ *for* $p > 2$. *This yields a conclusion that the rates of estimation in well-specified models (minimax risk) and in misspecified models (minimax regret) are equivalent in the regime* $p \in (0, 2]$. *In other words, for* $p \in (0, 2]$ *the problem of statistical learning enjoys the same minimax rate as the problem of statistical estimation. Our oracle inequalities also imply the* $\log(n)/n$ *rates for Vapnik-Chervonenkis type classes without the usual convexity assumption on the class; we show that these rates are optimal. As another corollary we obtain optimal rates of s-sparse convex aggregation. Finally, we introduce a more general risk measure that realizes a smooth transition between the minimax risk and the minimax regret depending on the magnitude of the approximation error. The minimax risk and the minimax regret appear as the two extremities of this scale. We provide sharp oracle inequalities for this new risk measure.* [Joint work with Alexander Rakhlin and Karthik Sridharan]

15:00 - 15:30 Coffee break

15:30 - 16:00 Sebastien Bubeck (Princeton University, USA) - Lipschitz Optimization with Noisy 0th Order Information

I will present an algorithm to optimize Lipschitz functions when only noisy 0th order information is available. The algorithm relies on a tree based structure and is called HOO (Hierarchical Optimistic Optimiz ation). The analysis depends on the 'near optimality dimension' of the function which measures the size of the set of near optimal points.

16:00 - 16:30 Vianney Perchet (Université Paris 7, France) - Generalized Exponential Weight Algorithm and Applications to Online Learning

I will show how the celebrated "exponential weight algorithm" can be generalized into the vectorial

setting (related to multi-criteria optimization) called Blackwell approachability. As applications, I will show how it can be used to construct simple and efficient algorithms that minimize refined versions of regret or that are calibrated, with respect to the family of all balls.

Thursday, June 27/2013

10:00 - 10:30 Boaz Nadler (Weizmann Institute, Israel) - On Estimation of Sparse Eigenvectors in High Dimensions

In this talk we'll discuss estimation of the population eigenvectors from a high dimensional sample covariance matrix, under a low-rank spiked model whose eigenvectors are assumed to be sparse. We present several models of sparsity, corresponding minimax rates and a procedure that attains these rates. We'll also discuss some differences between L_0 and L_q sparsity for $q > 0$, as well as some limitations of recently suggested SDP procedures.

10:30 - 11:00 Anatoli Juditsky (Université Joseph Fourier, France) - Nonparametric testing by convex optimization

We discuss a general approach to handling a class of nonparametric detection problems when the null and each particular alternative hypothesis states that the vector of parameters identifying the distribution of observations belongs to a convex compact. Our central result is a test for a pair of hypotheses of the outlined type which, under appropriate assumptions, is provably nearly optimal. The test is yielded by a solution to a convex programming problem, and, as a result, the proposed construction admits a computationally efficient implementation. We show how our approach can be applied to a rather general detection problem encompassing several classical statistical settings such as detection of abrupt signal changes, cusp detection and multi-sensor detection. [Joint work with Alexander Goldenshluger and Arkadi Nemirovski]

11.00 - 11:30 Coffee break

11:30 - 12:00 Olivier Catoni (Ecole Normale Supérieure, France) - Dimension Free PAC-Bayes Bounds and Unsupervised Classification

In this talk, I will present some dimension free PAC Bayes bounds and explain how they can be used to perform Principle Component Analysis and unsupervised clustering in high dimension.

12:00 - 12:30 Vladimir Spokoiny (WIAS, Germany and PremoLab MIPT, Russia) - Robust clustering using adaptive weights

This talk presents a new method of clustering in high dimension which is stable against outliers and does not require to know the number of clusters or their shape. Clustering structure of the data is described by a $n \times n$ matrix of weights whose elements are iteratively updates. The parameters of the method are calibrated by the "propagation" conditions and the results describe a separation distance between clusters which ensures cluster identification with a high probability. Numerical complexity of the method is of order $d \times n^2$ and it is feasible even in high dimension. The performance of the method is illustrated by the simulation study and a practical example. [Joint work with Elmar Diederichs (MIPT, Moscow)]

12:30 - 14:00 Lunch

14:00 - 14:30 Oleg Lepski (Université de Provence, France) - A New Method in Adaptive Estimation

During the talk I present the estimation procedure which is based on the selection from the family of kernel estimators those bandwidths are multivariate functions. The main ingredient of this method is the upper function for the L_p-norm of Gaussian processes, which, in its turn, has an independent interest.

14:30 - 15:00 Christophe Pouet (Université de Provence, France) - Classification of Sparse High-dimensional Vectors

We consider a classification problem with high-dimensional vector samples. We observe M samples drawn from M populations and we want to classify a new vector Z. We suppose that the difference between the distributions of the populations is only in a shift that is a sparse vector. We obtain asymptotically (as the dimension d tends to infinity) sharp classification boundary for the Gaussian noise and fixed sample size, and we propose classifiers that provide this boundary. [Joint work with Yuri Ingster]

15:00 - 15:30 Coffee break

15:30 - 16:00 Arnak Dalalyan (ENSAE - CREST, France) - Sparse Regression under Heteroskedasticity and Group Sparsity

Popular sparse estimation methods based on ℓ_1 relaxation, such as the Lasso and the Dantzig selector, require the knowledge of the variance of the noise in order to properly tune the regularization parameter. This constitutes a major obstacle in applying these methods in several frameworks such as time series, random fields, inverse problems for which the noise is rarely homoscedastic and its level is hard to know in advance. In this paper, we propose a new approach to the joint estimation of the conditional mean and the conditional variance in a high dimensional (auto) regression setting. An attractive feature of the proposed estimator is that it is efficiently computable even for very large scale problems by solving a second order cone program (SOCP). We present theoretical analysis and numerical results assessing the performance of the proposed procedure.

16:00 - 16:30 Laurent Cavalier (Université de Provence, France) - Model Selection in Sparse Heterogeneous Framework

We consider a Gaussian sequence space model $X_\lambda = f_\lambda + \xi_\lambda$, where ξ has a diagonal covariance matrix $\Sigma = \mathrm{diag}(\sigma_\lambda^2)$. This heterogeneous model may appear in frameworks where the variance is fluctuating, for example in inverse problems or fractional Brownian motion. We consider mainly

the situation where the parameter is sparse. Our goal is to estimate the unknown parameter by a model selection approach. The heterogenous case is much more involved than the direct model. Indeed, there is no more symmetry inside the stochastic process that one needs to control, since each empirical coefficient has its own variance. The problem and the penalty do not only depend on the number of coefficients that one selects, but also on their position. This appears also in the minimax bounds where the worst coefficients will go to the larger variances. However, with a careful and explicit choice of the penalty we are able to select the correct coefficients and get a sharp non-asymptotic control of the risk of our procedure. Results are also obtained for full model selection and a family of thresholds. We obtain a minimax upper bound, the estimator almost attains the lower bound (up to a constant 2). Moreover, the procedure is fully adaptive, we obtain an explicit penalty, valid in the mathematical proofs and in simulations. [Joint work with Markus Reiss]

16:30 - 17:00 David Gamarnik (MIT, USA) - Limits of Local Algorithms for Sparse Random Graphs

Algorithms which can be run in parallel on large networks based on local information (local algorithms) gained a lot of prominence recently in a variety of applications, primarily as a way of addressing the scaling issues for computations on large instances. Some local algorithms such as, for example, the Belief Propagation algorithm emerged as strong contenders for solving a variety of optimization and inference problems on large scale network models, including random instances of hard constraint satisfaction problems. In this talk we will discuss limitations of local algorithms for solving constraint satisfaction problems on random graphs. In particular, we will discuss a negative resolution of a recent conjecture regarding the power of local algorithms for solving largest independent set problem on random graphs. [Joint work with Madhu Sudan]

Friday, June 28/2013

10:00 - 10:30 Vladimir Vyugin (Institute for Information Transmission Problems, Russia) - An Application of the Foster - Vohra Method of Calibration to Stock Market Games

A new application of the Foster–Vohra method of calibration to stock market games will be presented. We present a universal method for algorithmic trading in Stock Market which performs asymptotically at least as well as any stationary trading strategy that computes the investment at each step using a fixed function of the side information that belongs to a given RKHS (Reproducing Kernel Hilbert Space). Using a universal kernel, we extend this result for any continuous stationary strategy. In this learning process, a trader chooses his gambles using predictions made by a randomized well-calibrated algorithm. Our strategy is based on Dawid's notion of calibration with more general checking rules and on some modification of Kakade and Foster's randomized rounding algorithm for computing the well-calibrated forecasts. We combine the method of randomized calibration with Vovk's method of defensive forecasting in RKHS. Unlike in statistical theory, no stochastic assumptions are made about the stock prices. Our empirical results on historical markets provide strong evidence that this type of technical trading can "beat" some generally accepted trading strategies if transaction costs are ignored.

10:30 - 11:00 Konstantin Vorontsov (Moscow State University, Russia) - Combinatorial Theory of Overfitting

Overfitting is one of the most challenging problems in Statistical Learning Theory. Classical approaches recommend to restrict complexity of the search space of classifiers. Recent approaches benefit from more refined analysis of a localized part of the search space. Combinatorial theory of overfitting is a new developing approach that gives tight data dependent bounds on the probability of overfitting. It requires a binary loss function and uses a detailed representation of the search space in a form of a directed acyclic graph. The size of the graph is usually enormous, however the bound can be effectively estimated by walking through its small localized part that contains best classifiers. We consider exact combinatorial bounds for some nontrivial model sets of classifiers. Also we apply combinatorial bounds on real data sets to build voting ensembles of low dimensional linear classifiers and conjunction rules.

11.00 - 11:30 Coffee break

11:30 - 12:00 Yury Yanovich (Institute for Information Transmission Problems and Premolab, Russia) - Asymptotically Optimal Method for Manifold Estimation Problem

Manifold learning is considered as manifold estimation problem: to estimate an unknown well-conditioned q-dimensional manifold embedded in a high-dimensional observation space given sample of n data points from the manifold. It is shown that the proposed Grassmann & Stiefel Eigenmaps algorithm estimates the manifold with a rate n to the power of $-2/(q+2)$, where

q is dimension of the manifold; this rate coincides with a minimax lower bound for Hausdorff distance between the manifold and its estimator (*Genovese et al. Minimax manifold estimation. Journal of machine learning research, 13, 2012).* [Joint work with Alexander Kuleshov and Alexander Bernstein (IITP and PreMoLab, Moscow)]

12:00 - 12:30 Karim Lounici (Georgia Tech, USA) - Estimation and Variable Selection with Exponential Weights

In the context of a linear model with a sparse coefficient vector, exponential weights methods have been shown to be achieve oracle inequalities for prediction. We show that such methods also succeed at variable selection and estimation under the near minimum condition on the design matrix, instead of much stronger assumptions required by other methods such as the Lasso or the Dantzig Selector. The same analysis yields consistency results for Bayesian methods and BIC type variable selection under similar conditions. [Joint work with Ery Arias Castro]