FOCM 2014 - Workshop C3
Learning Theory

C3 - December 18, 14:30 – 15:00

A Tale of Three Regression Problems

Alexander Rakhlin
University of Pennsylvania, USA
rakhlin@wharton.upenn.edu

We consider the problem of regression in three scenarios: (a) random design under the assumption that the model F is correctly specified, (b) distribution-free statistical learning with respect to a reference class F; and (c) online regression with no assumption on the generative process. The first problem is often studied in the literature on nonparametric estimation, the second falls within the purview of statistical learning theory, and the third is studied within the online learning community. Do these three problems really differ from the minimax point of view? This question will be addressed in this talk.

Joint work with Karthik Sridharan (Cornell University) and Sasha Tsybakov (ENSAE-Paris Tech).

C3 - December 18, 15:00 – 15:30

Stochastic proximal methods for online learning

Silvia Villa
Istituto Italiano di Tecnologia, Italia
silvia.villa@iit.it

In this talk I will present recent advances on the convergence properties of a class of stochas- tic proximal gradient algorithms for solving minimization problems. These algorithms are easy to implement and suitable for solving high dimensional problems thanks to the low memory requirement of each iteration. Moreover, they are particularly suitable for compos- ite optimization, where a convex objective function is the sum of a smooth and a non-smooth component. I will show that this algorithm can be naturally applied to solve standard online machine learning algorithms and I will focus on convergence in expectation and convergence almost surely of the iterates.

C3 - December 18, 15:35 – 16:25

Efficient minimax strategies for online prediction

Peter Bartlett
UC Berkeley and Queensland University of Technology, USA
peter@berkeley.edu

Consider prediction games in which, in each round, a strategy makes a decision, then observes an outcome and pays a loss. The aim is to minimize the regret, which is the amount by which the total loss incurred exceeds the total loss of the best decision in hindsight. We study the case where decisions and outcomes lie in a convex subset of a Hilbert space, and loss is squared distance. When the set is the simplex, this is

the 'Brier game,' studied for the calibration of sequential probability forecasts; when it is the Euclidean ball, the game is related to sequential Gaussian density estimation. We show that the value of the game depends only on the radius of the smallest ball that contains the convex subset, and that the minimax optimal strategy is a simple shrinkage strategy that can be efficiently computed, given the center of the smallest ball.

Joint work with Wouter Koolen (UC Berkeley and Queensland University of Technology) and Alan Malek (UC Berkeley).

C3 - December 18, 17:00 – 17:30

Simultaneous Model Selection and Learning through Parameter-free Stochastic Gradient Descent

Francesco Orabona
Yahoo! Labs NY, USA
francesco@orabona.com

Stochastic gradient descent algorithms for training linear and kernel predictors are gaining more and more importance, thanks to their scalability. While various methods have been proposed to speed up their convergence, the issue of the model selection phase has often been ignored in the literature. In fact, in theoretical works most of the time unrealistic assumptions are made, for example, on the prior knowledge of the norm of the optimal solution. Hence, costly validation methods remain the only viable approach in practical applications. In this talk, we show how a family of kernel-based stochastic gradient descent algorithms can perform model selection while training, with no parameters to tune, nor any form of cross-validation, and only one pass over the data. These algorithms are based on recent advancements in online learning theory in unconstrained settings. Optimal rates of convergence will be shown under standard smoothness assumptions on the target function, as well as empirical results.

C3 - December 18, 17:30 – 18:00

Trivial Pursuit: A Shallow Learning Retrospective

Benjamin Recht
University of California, Berkeley, USA
brecht@eecs.berkeley.edu

In this talk, I review the history of random features in machine learning over the last ten years. Presenting both my work on the subject and a survey of the related literature, I will show how shallow banks of random features are able to match the performance of complicated learned or engineered features on a variety of difficult learning tasks. I will then describe the present challenges in random feature design and our progress towards making these features scalable, parsimonious, and interpretable.

C3 - December 18, 18:00 – 18:30

The Wasserstein Barycenter Problem: Formulation, Computation and Applications

Marco Cuturi
Kyoto University, Japan
mcuturi@i.kyoto-u.ac.jp

How can we define the average of a set of probability measures? This question is important because (1) averaging ranks among the most elementary operations used in statistics to summarize data (2) probability measures are ubiquitous in machine learning, where they are used to represent datasets, generative models or complex objects (an image as a bag-of-features, a text as a bag-of-words).

I will present in this talk a possible answer to this question grounded on the optimal transport (a.k.a. Wasserstein/ earth mover's) geometry. The problem I will describe, known as the Wasserstein barycenter problem, tries to find, given a set of probability measures of interest, the probability measure that minimizes the sum of all its Wasserstein-distances to those probability measures. After providing a few self-contained reminders on optimal transport in the first part of the talk, I will illustrate using toy data that Wasserstein barycenters have several intuitive and appealing properties. I will then show that in its original form theWasserstein barycenter problem is intractable, but that it can be solved approximately, very efficiently, and to arbitrary precision in practice by regularizing it with an entropic term. I will provide details of very recent algorithmic advances in this nascent field followed by an application to the visualization of datasets of brain activation maps.

C3 - December 19, 14:30 – 15:00

Learning a Hidden Basis through Imperfect Measurements: Why and How

Misha Belkin
Ohio State University, USA
mbelkin@cse.ohio-state.edu

In this talk I will describe a general framework of inferring a hidden basis from imperfect measurements. I will show that a number of problems from the classical eigendecompositions of symmetric matrices to such topics of recent interest as multiclass spectral clustering, Independent Component Analysis and Gaussian mixture learning can be viewed as examples of hidden basis learning.

I will then describe algorithms for basis recovery and provide theoretical guarantees in terms of computational complexity and perturbation size. The proposed algorithms are based on what may be called "gradient iteration" and are simple to describe and to implement. They can be viewed as generalizations of both the classical power method for recovering eigenvectors of symmetric matrices as well as the recent work on power methods for tensors. Unlike these methods, our analysis is based not on tensorial properties, but on certain "hidden convexity" of constrast functions.

Joint work with L. Rademacher (Ohio State University) and J. Voss (Ohio State University).

C3 - December 19, 15:00 – 15:30

Tensor decomposition, convex optimization, and multitask learning

Ryota Tomioka
Toyota Technological Institute at Chicago, USA
tomioka@ttic.edu

Tensor factorization, or multilinear modelling, has received much attention recently. Compared to its two-dimensional counterpart, matrix factorization, many properties related to tensors, for example, the rank, are known to be hard to compute. Recently new approaches based on convex relaxation of tensor (multilinear-)rank have emerged. Although, these new methods come with worst case performance guarantees, they tend to be less efficient than previously known greedy algorithms in practice. I will overview and

discuss the possibility and limitation of these approaches from the perspective of computation-statistics trade-off. Furthermore, I will present a recent application of the above idea to multi-task learning.

Joint work with Kishan Wimalawarne (Tokyo Institute of Technology,), Taiji Suzuki (Tokyo Institute of Technology), Kohei Hayashi (National Institute of Informatics, Tokyo), Hisashi Kashima (Kyoto University) and Masashi Sugiyama (University of Tokyo).

C3 - December 19, 15:30 – 16:00

Democratic Learning: Learning to Represent Data for Everybody

Guillermo Sapiro
Duke University, USA
guillermo.sapiro@duke.edu

In this talk I will describe a simple framework for learning data transforms that are computationally free and help diverse classification and clustering algorithms. When incorporated into standard techniques such as subspace clustering, random forests, and hashing codes, we obtain one to two orders of magnitude improvement at virtually no cost. I will present both the underlying concepts and applications ranging from scene recognition to image classification to 3D object analysis.

Joint work with Qiang Qiu (Duke University) and Alex Bronstein (Tel Aviv University).

C3 - December 19, 16:00 – 16:30

Stability and statistical properties of topological information inferred from data

Frederic Chazal
INRIA Saclay, France
frederic.chazal@inria.fr

Computational topology has recently seen an important development toward data analysis, giving birth to Topological Data Analysis. Persistent homology appears as a fundamental tool in this field. It is usually computed from filtrations built on top of data sets sampled from some unknown (metric) space, providing "topological signatures" revealing the structure of the underlying space. In this talk we will present a few stability and statistical properties of persistence diagrams that allow to efficiently compute relevant topological signatures that are robust to the presence of outliers in the data.

C3 - December 19, 17:00 – 17:30

Algebraic Combinatorial Single-Entry Low-Rank Matrix Completion

Franz Kiraly
University College London, UK
f.kiraly@ucl.ac.uk

In recent years, the low-rank matrix completion model has enjoyed quite some success for recommendation and prediction learning. Many standard algorithms in the field are designed for completing the whole matrix (or derivates) - which also means that achieving scalability on huge data sets is a challenging task.

The algebraic combinatorial approach aims for scalability in a canonical and non-heuristic way: instead of considering "global" properties of the matrix - e.g. low nuclear norm, sparse factorization - the underlying theory describes low-rankness "locally", in terms of algebraic relations between small numbers of entries and combinatorial properties of the observation pattern, both closely interrelated. Algorithmically, this allows to obtain estimates for single entries and error bounds while looking only at a small part of the observation data in the "combinatorial neighbourhood", described by the bipartite graph of observations.

The algebraic combinatorial approach therefore allows, for the first time, a systematic treatment of single-entry estimates including single-entry error bounds, and it yields, for the first time, a closed approach to the low-rank model that is intrinsically local.

In the talk, I will give a brief introduction to the matrix completion problem and its algebraic combinatorial formulation; I will demonstrate how this allows to derive simple reconstruction algorithms, and review some recent empirical results.

Joint work with Duncan Blythe (TU Berlin), Louis Theran (Aalto University, Helsinki) and Ryota Tomioka (TTI Chicago).

C3 - December 19, 17:30 – 18:00

Neurally Plausible Algorithms Find Globally Optimal Sparse Codes

Ankur Moitra
Massachusetts Institute of Technology, USA
moitra@mit.edu

We prove that neurally plausible algorithms — including the classic one identified by Olshausen and Field — can efficiently find a basis that enables sparse representation of a dataset, a foundational problem in neural computation and machine learning. This problem involves non-convex optimization. However, under plausible conditions where the global optimum is unique, we show that the algorithms converge rapidly and with near-optimal sample complexity, to the global optimum. This suggests that non-convexity need not be a hurdle to a rigorous mathematical and algorithmic theory of neural computation.

Joint work with Sanjeev Arora (Princeton University), Rong Ge (Microsoft Research) and Tengyu Ma (Princeton University).

C3 - December 20, 14:35 – 15:25

Kernel-based learning methods

Ingo Steinwart
University of Stuttgart, Germany
ingo.steinwart@mathematik.uni-stuttgart.de

The last decade has witnessed an explosion of data collected from various sources. Since in many cases these sources do not obey the assumptions of classical statistical approaches, new automated methods for interpreting such data have been developed in the machine learning community. Statistical learning theory tries to understand the statistical principles and mechanisms these methods are based on.

This talk begins by introducing some central questions considered in statistical learning. Then various theoretical aspects of a popular class of learning algorithms, which include support vector machines, are discussed. In particular, I will describe how classical concepts from approximation theory such as

interpolation spaces and entropy numbers are used in the analysis of these methods. The last part of the talk considers more practical aspects including the choice of the involved loss function and some implementation strategies. In addition, I will present a data splitting strategy that enjoys the same theoretical guarantees as the standard approach but reduces the training time significantly.

C3 - December 20, 15:30 – 16:00

Sparse Estimation with Strongly Correlated Variables

Robert Nowak
University of Wisconsin-Madison, USA
nowak@ece.wisc.edu

This talk considers ordered weighted L1 (OWL) norm regularization for sparse estimation problems with strongly correlated variables. We show that OWL norm regularization automatically clusters strongly correlated variables, in the sense that the coefficients associated with such variables have equal estimated values. Furthermore, we characterize the statistical performance of OWL norm regularization for generative models in which certain clusters of regression variables are strongly (even perfectly) correlated, but variables in different clusters are uncorrelated. We show that if the true p-dimensional signal generating the data involves only s of the clusters, then O(s log p) samples suffice to accurately estimate the signal, regardless of the number of coefficients within the clusters. The estimation of s-sparse signals with completely independent variables requires just as many measurements. In other words, using the OWL we pay no price (in terms of the number of measurements) for the presence of strongly correlated variables.

C3 - December 20, 16:00 – 16:30

Learning Theory and Adaptive Partitioning in High Dimensions

Peter Binev
University of South Carolina, USA
binev@mailbox.sc.edu

Adaptive partitioning has been one of the methods of choice for several problems from nonlinear approximation theory. A typical challenge in applying this approach to learning theory is the increased complexity of the high-dimensional realization of the adaptive algorithms. We discuss paradigms like sparse occupancy and decorated trees that are designed to alleviate the difficulties related to high dimensions and tuned to certain learning theory setups.

C3 - December 20, 17:30 – 18:00

Iterative Regularization for Computational Learning

Lorenzo Rosasco
Universita' di Genova , Italy
lrosasco@mit.edu

Iterative regularization approaches to ill-posed inverse problems are known to provide a viable alternative to Tikhonov regularization, especially in large scale problems. Supervised learning can be seen as an inverse problem under a suitable stochastic data model. In this context, iterative regularization is

particularly suited since statistical and computational aspects are tackled at once, a key property when dealing with large data-sets. In this talk we will discuss old and new results on learning with iterative regularization and connect them with recent results in online learning.
