Francis Bach (DI ENS, Centre de recherche INRIA PARIS; France)

Title: Statistical Optimality of Stochastic Gradient Descent on Hard Learning Problems through Multiple Passes

Abstract: We consider stochastic gradient descent (SGD) for least-squares regression with potentially several passes over the data. While several passes have been widely reported to perform practically better in terms of predictive performance on unseen data, the existing theoretical analysis of SGD suggests that a single pass is statistically optimal. While this is true for low-dimensional easy problems, we show that for hard problems, multiple passes lead to statistically optimal predictions while single pass does not; we also show that in these hard models, the optimal number of passes over the data increases with sample size. In order to define the notion of hardness and show that our predictive performances are optimal, we consider potentially infinite-dimensional models and notions typically associated to kernel methods, namely, the decay of eigenvalues of the covariance matrix of the features and the complexity of the optimal predictor as measured through the covariance matrix. We illustrate our results on synthetic experiments with non-linear kernel methods and on a classical benchmark with a linear model.

Joint work with Loucas Pillaud-Vivien and Alessandro Rudi.

Gérard Biau (LPSM, Sorbonne Université, Paris; France)

Title: Some theoretical properties of GANs

Abstract: Generative Adversarial Networks (GANs) are a class of generative algorithms that have been shown to produce state-of-the art samples, especially in the domain of image creation. The fundamental principle of GANs is to approximate the unknown distribution of a given data set by optimizing an objective function through an adversarial game between a family of generators and a family of discriminators.
In this presentation, we offer a better theoretical understanding of GANs by analyzing some of their mathematical and statistical properties.

We study the deep connection between the adversarial principle underlying GANs and the Jensen-Shannon divergence, together with some optimality characteristics of the problem. An analysis of the role of the discriminator family via approximation arguments is also provided. In addition, taking a statistical point of view, we study the large sample properties of the estimated distribution and prove in particular a central limit theorem. Some of our results are illustrated with simulated examples.

Joint work with B. Cadre (ENS Rennes), M. Sangnier (Sorbonne University), and U. Tanielian (Sorbonne University & Criteo)

Richard Combes (Laboratoire des signaux et systèmes, CentraleSupélec; Gif sur Yvette; France)

Title: Minimal Exploration in Structured Stochastic Bandits

Abstract: In this talk we introduce and address a wide class of stochastic bandit problems where the function mapping the arm to the corresponding reward exhibits some known structural properties. Most existing structures (e.g. linear, Lipschitz, unimodal, combinatorial, dueling, ...) are covered by our framework. We derive an asymptotic instance-specific regret lower bound for these problems, and develop OSSB, an algorithm whose regret matches this fundamental limit. OSSB is not based on the classical principle of "optimism in the face of uncertainty" or on Thompson sampling, and rather aims at matching the minimal exploration rates of sub-optimal arms as characterized in the derivation of the regret lower bound.

Camille Couprie (Facebook; France)

Title: Future video prediction and creative image generation)

Abstract: This presentation will deal with conditioned and unconditioned generative models.
An important prerequisite towards intelligent behavior is the ability to anticipate future events. Predicting the appearance of future video frames is a proxy task towards pursuing this ability. We will present how generative adversarial networks (GANs) can help, and novel approaches predicting in higher level feature spaces.

In a second part, we will see how to develop the abilities of GANs to deviate from training examples to generate creative images

Tim van Erven (Statistics group, Leiden University, Leiden; Netherlands)

Title: MetaGrad: Multiple learning rates in online learning

Abstract: In online convex optimization it is well known that certain subclasses of objective functions are much easier than arbitrary convex functions. We are interested in designing adaptive methods that can automatically get fast rates in as many such subclasses as possible, without any manual tuning. Previous adaptive methods are able to interpolate between strongly convex and general convex functions. We present a new method, MetaGrad, that adapts to a much broader class of functions, including exp-concave and strongly convex functions, but also various types of stochastic and non-stochastic functions without any curvature.
For instance, MetaGrad can achieve logarithmic regret on the unregularized hinge loss, even though it has no curvature, if the data come from a favourable probability distribution. MetaGrad's main feature is that it simultaneously considers multiple learning rates. Unlike all previous methods with provable regret guarantees, however, its learning rates are not monotonically decreasing over time and are not tuned based on a theoretically derived bound on the regret. Instead, they are weighted directly proportional to their empirical performance on the data using a tilted exponential weights master algorithm.

References:
* T. van Erven and W.M. Koolen. MetaGrad: Multiple Learning Rates in Online Learning. NIPS 2016.

* W.M.Koolen, P. Grünwald and T. van Erven. Combining Adversarial Guarantees and Stochastic Fast Rates in Online Learning. NIPS 2016.

Nicolas Gillis (Université de Mons; Belgium)

Title: Computing nonnegative matrix factorizations

Abstract: Given a nonnegative matrix M and a factorization rank r, nonnegative matrix factorization (NMF) is the problem of finding two nonnegative matrices, U with r columns and V with r rows, such that the product U*V approximates M. NMF has become a widely used tool for the analysis of high-dimensional data as it automatically extracts sparse and meaningful features from a set of nonnegative data vectors. We first illustrate this property of NMF on three applications, in image processing, text mining and hyperspectral imaging. Then we address the problem of computing NMFs and discuss the following aspects: (1) computational complexity issues, (2) standard non-linear optimization schemes and acceleration, (3) exact NMF, which requires M=UV, and its geometric interpretation, and (4) separable NMF that can be solved efficiently, even in the presence of noise, under an appropriate assumption on the input matrix M with applications in text mining and hyperspectral imaging.

Emmanuel Gobet (CMAP, Ecole Polytechnique, Orsay; France)

TBA

José Miguel Hernández-Lobato (Dpt of Engineering, University of Cambridge; UK)

Title: Bayesian optimization for accelerated exploration of chemical space

Abstract: Chemical space is so large as to make a brute force search for molecules with improved properties infeasible. Bayesian optimization methods can accelerate the discovery process by sequentially identifying the most useful experiments to be performed next. However, existing methods have shortcomings that limit their applicability to the molecule search problem. First, they lack scalability to the large amounts of data that are required to successfully navigate chemical space. Second, they are unable to learn feature representations for the data, which reduces their statistical efficiency in the large data scenario. Third, they often fail when the search space is discrete as is the case of chemical space. In this talk I will give a brief introduction to Bayesian optimization methods and then I will present different contributions that aim to solve or at least alleviate the aforementioned problems.

Anatoli Juditsky (Laboratoire J. Kuntzmann, Université Grenoble Alpes, France)

TBA

Andreas Krause (Leaning and Adaptive Systems group, ETH Zurich; Switzerland)

Title: Bayesian optimisation and Gaussian process bandits: Theory and Applications

Abstract: Bayesian optimisation (BO) is an approach towards global optimisation of functions that are accessible only via a noisy black box that may be expensive to evaluate. It has found numerous applications ranging from calibrating computer simulations to automatic hyperparameter tuning of machine learning models to policy search in reinforcement learning. The key idea is to probabilistically model the objective function — often with Gaussian processes — and use the uncertainty in the model to trade exploration and exploitation. In this tutorial I will provide an introduction to Bayesian optimisation with Gaussian processes and discuss formal connections to multi-armed bandits with RKHS payoff functions. These connections led to the design of algorithms for BO with strong theoretical guarantees. Beyond discussing basic approaches and the key ideas behind their analysis, I will present natural extensions to BO with side information, multi-objective and multi-fidelity BO, exploiting parallelism as well as recent work on safe exploration. In addition, I will discuss several applications ranging from optimal experimental design of wet lab experiments to robotic control.

Jeorges Lan (Georgia Institute of Technology, Atlanta; USA)

Title: Stochastic Optimization Algorithms for Machine Learning.

Abstract: Stochastic optimization plays a fundamental role in modern machine learning. In this short course, we introduce different types of stochastic optimization algorithms that have been widely used in training machine learning models, including stochastic gradient descent (SGD), stochastic mirror-descent, accelerated stochastic gradient descent (SGD with momentum), nonconvex SGD and its acceleration, variance reduction techniques, randomized incremental gradient methods, and distributed stochastic algorithms. We focus on the complexity issues associated with these algorithms. Some practical enhancements will be briefly discussed.

Jason Lee (Data Sciences and Operations Dpt, Univ. Southern California; USA)

Title: Geometry of Optimization Landscapes and Implicit Regularization of Optimization Algorithms.

Abstract: We first study the problem of learning a Gaussian input two-layer ReLU network with positive output layer and and the symmetric matrix completion problem. Despite the non-convexity of both problems, we prove that every local minimizer is a global minimizer. Since gradient descent converges to local minimizers, this shows that simple gradient-based methods can find the global optimum of these non-convex problems.

In the second part, we analyze the implicit regularization effects of various optimization algorithms. In particular we prove that for least squares with mirror descent, the algorithm converges to the closest solution in terms of the bregman divergence. For linearly separable classification problems, we prove that the steepest descent with respect to a norm solves SVM with respect to the same norm. For over-parametrized non-convex problems such as matrix sensing or neural net with quadratic activation, we prove that gradient descent converges to the minimum nuclear norm solution, which allows for both meaningful optimization and generalization guarantees.

This is joint work with Rong Ge, Suriya Gunasekar, Tengyu Ma, Mor Shpigel, Daniel Soudry, and Nati Srebro.

Fabien Panloup (LAREMA, Université d'Angers; France)

Title: Non asymptotic analysis of the Ruppert Polyak averaging algorithm

Abstract: This talk, based on a joint work with Sébastien Gadat, is devoted to the non-asymptotic control of the mean-squared error for the Ruppert-Polyak stochastic averaged gradient descent. Since the paper by Polyak and Juditsky, the algorithmis known to be optimal from an asymptotic point of view. In the non-asymptotic
setting, some results exist in a uniformly convex setting or in some particular pathological cases.
Here, we will present some new results in this topic, beginning by a general theorem, which provides some general conditions which ensure non-asymptotic tight bounds (optimal with respect to the Cramer-Rao lower bound). Then, we will particularly develop applications in some non-uniformly convex settings under a so-called Kurdyka-Lojasiewicz-type condition. This general condition allows us to recover some important examples such as on-line learning for logistic regression and recursive quantile estimation.

Csaba Szepesvári (Deepmind; UK)

Title: Completing the classification of adversarial partial monitoring games)

Abstract: Partial monitoring is a generalization of bandit problems where the relationship between the feedback and the loss is loosened. As such, partial monitoring offers a rich framework to study the exploration-exploitation dilemma.

In the finite, adversarial version of the problem that we consider here a learner and an adversary take actions in a sequential fashion from their respective finite action sets. A pair of actions result in a fixed loss and a fixed observation; the mapping that maps pairs of actions to the losses and observations is given to the learner ahead of time. The unknown are the actions taken by the adversary. Rather, the learner gains information of the adversaries actions by receiving the observation underlying the joint action of a round. Since what observations can be received in a round is governed by the action taken, the learner has control over what it can receive information about. The learner's goal, as usual, is to keep its total regret, the difference between the total loss suffered and the loss that could have been suffered had the had full knowledge of the adversaries actions and played the best response to these, as small possible.

The classification problem of partial monitoring games is concerned with determining how to play in a given game described by the map from joint actions to losses and observations so as to achieve at most a constant multiple of the minimax regret. The study of partial monitoring games started with the work Rusitchini in 1999, and significant advances were made by Piccolboni and Schindelhauer, Cesa-Bianchi, Lugosi and and Stoltz. Around 2010, the goal of the full classification emerged and an almost complete answer was given by the joint work of Bartok, Foster, Pal, Rakhlin and Szepesvari (2014).

This work gave an almost complete characterization of all possible partial monitoring and identified four regimes: trivial, easy, hard and impossible games. The characterization, however, left out a set of games, namely those when there is at least one actions that are optimal under some play of the adversary, but the region where this happens has a dimension defect. As a result, even though these "tricky" actions are

nondominated, they could be omitted if not for the information they bring in. This creates a tricky situation when designing algorithms and until now deciding whether the presence of these tricky actions can move a game from the easy to the hard category remained open. The answer was known to be "no" for stochastic games when the adversary plays a fixed stochastic strategy. In this talk, I present the solution to this problem: It turns out that a simplified and improved version of the algorithm by Foster and Rakhlin can achieve O(T^(1/2)) regret for these problems, thus resolving the open problem and completing the characterization of finite adversarial partial monitoring games. In the talk, I will cover the key ideas of the characterization in previous works, then motivate and explain the new ideas that led to the new algorithm and explain the proof of the new result.

The talk is based on joint work with Tor Lattimore.

Ohad Shamir (Dpt Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot; Israël)

Title: Optimization Landscape of Neural Networks: Where Do the Local Minima Hide?

Abstract: Training neural networks is a highly non-convex optimization problem, which is often successfully solved in practice, but the reasons for this are poorly understood. Much recent work has focused on showing that these non-convex problems do not suffer from poor local minima. However, this has only been provably shown under strong assumptions or in highly restrictive settings. In this talk, I'll describe some recent results on this topic, both positive and negative. On the negative side, I'll show how local minima can be ubiquitous even when optimizing simple, one-hidden-layer networks, under the nicest possible data distributions. On the flip side, I'll discuss how looking at other architectures (such as residual units), or modifying the question, can lead to positive results under mild assumptions.

Includes joint work with Itay Safran and Yossi Arjevani.

Michal Valko (SequeL team, INRIA Lille - Nord Europe; Lille; France)

Title: Active block-matrix completion with adaptive confidence sets

Abstract: We address the problem of an active setting for a matrix completion, where the learner can choose, from which sub-matrix, it receives a sample (drawn uniformly at random). Our main practical motivation is the market segmentation, where the blocks are different regions with different preferences of the customers. The challenge in this setting is that each of the submatrix can be of a different size and also of a different rank. We provide and analyze a new algorithm, MAlocate that is able to adapt to the ranks of the different sub-matrices. We also prove a lower-bound showing that our strategy is minimax-optimal, and we demonstrate its performance with experiments.

Join work with Andrea Locatelli and Alexandra Carpentier

René Vidal (Johns Hopkins University, USA)

Title: Global Optimality in Matrix Factorization, Tensor Factorization and Deep Learning

Abstract: The past few years have seen a dramatic increase in the performance of recognition systems thanks to the introduction of deep networks for representation learning. However, the mathematical reasons for this success remain elusive. A key issue is that the neural network training problem is non-convex, hence optimization algorithms may not return a global minima. In addition, the regularization properties of algorithms such as dropout remain poorly understood. Building on ideas from convex relaxations of matrix factorizations, this work proposes a general framework which allows for the analysis of a wide range of non-convex factorization problems – including matrix factorization, tensor factorization, and deep neural network training. The talk will describe sufficient conditions under which a local minimum of the non-convex optimization problem is a global minimum and show that if the size of the factorized variables is large enough then from any initialization it is possible to find a global minimizer using a local descent algorithm. The talk will also present an analysis of the optimization and regularization properties of dropout in the case of matrix factorization.

Lenka Zdeborova (Institut de Physique Théorique, CEA Saclay; Paris; France)

Title: Constrained low-rank matrix completion

Abstract: Low-rank matrix factorization is one of the basic methods used in data analysis for unsupervised learning of relevant features and other types of dimensionality reduction. We consider a probabilistic model of constrained low-rank matrix (or tensor) estimation where the factors are drawn uniformly at random and the low-rank matrix (or tensor) is observed through a general component0wise output channel. This is a generalization of the popular spiked covariance model with iid spikes. We present a generic methodology coming from statistical physics that leads to a closed formula for the minimum-mean-squared error achievable in this model by the Bayes-optimal estimator. We also present the corresponding approximate message passing algorithms and locate a region of parameters for which this algorithms achieves the optimal performance. We discuss intuition on computational hardness of the complementary region. Our analysis also provides results and insight on performance of commonly used spectral algorithms.