IEEE EUROPEAN SCHOOL OF INFORMATION THEORY 2023
MONDAY 17TH JULY – FRIDAY 21ST JULY

4-hour Tutorial speakers:

Tobias Koch

Title: From the law of large numbers to the saddlepoint expansion and their application to channel coding theorems

Abstract: Convergence results on the distribution of the average of random variables are key ingredients in the proofs of channel coding theorems. For example, the direct part of Shannon's channel coding theorem is based on the Asymptotic Equipartition Property, which is a direct consequence of the law of large numbers. Error exponents follow from the theory of large deviations. And the normal approximation rests upon the central limit theorem. In the first part of this tutorial, I will review important convergence results, such as the law of large numbers, Cramer's theorem, and the Berry-Esseen theorem. I then show how these results can be combined to obtain the saddlepoint expansion. In the second part of the tutorial, I will discuss how these convergence results can be applied to derive channel coding theorems. I then argue that these theorems may give rise to accurate performance benchmarks for communication channels. In particular, the saddlepoint expansion appears to be very precise for a large range of blocklengths and error probabilities.

Rashmi Vinayak

Title: Coding theory for distributed systems

Abstract: Distributed systems often operate in regimes prone to failures, unavailabilities, and high variance in performance. Redundancy is a common approach employed to impart resilience and mitigate these adverse effects. Given the tight constraints on resources, this brings up the critical challenge of introducing resource-efficient redundancy at various levels of the system stack. Coding theory provides many valuable tools to address this challenge, and these tools are increasingly finding their way into modern computer systems. This tutorial will present coding theory tools for distributed data storage systems and distributed computation systems. The topics will cover both theoretical results as well as practical aspects, including examples of real-world deployments.

Mary Wootters

Title: List-Decoding and List-Recovery in Error Correcting Codes

Abstract: The goal of an error-correcting code is to allow a receiver to reliably recover a message, even in the presence of noise. In list-decoding, and the related notion of list-recovery, the goal is relaxed so that the receiver is allowed to return a short list of possible messages, but in a regime where there is so much noise that unique decoding is impossible. Beyond communication, list-decoding and list-recovery have applications in algorithm design, complexity theory, mathematics, and elsewhere in coding theory. In this tutorial, we will cover recent research in this area, including constructions and known impossibility results, as well as applications and open questions.

Spotlight talks (2 hours):

András György (Talk 1).

Title: Transfer Learning with Pretrained Classifiers

Abstract: We study the ability of foundation models to learn representations for classification that are transferable to new, unseen classes. Recent results in the literature show that representations learned by a single classifier over many classes are competitive on few-shot learning problems with representations learned by special-purpose algorithms designed for such problems. We offer a theoretical explanation for this behavior based on the recently discovered phenomenon of class-feature-variability collapse, that is, that during the training of deep classification networks the feature embeddings of samples belonging to the same class tend to concentrate around their class means.

András György (Talk 2):

Title: Zeroth order (stochastic) convex optimization

Abstract: This talk will review general approaches to zeroth order (stochastic) convex optimization including some recent advances.

Varun Jog:

Title: The sample complexity of simple binary hypothesis testing

Abstract: Let p and q be two distributions over a finite domain, and let X_1, ..., X_n be i.i.d. samples drawn from either p or q. The goal of simple binary hypothesis testing is to determine whether the samples were drawn from p or q. This problem can be formulated with or without a prior distribution on the choice of p or q. The sample complexity of this problem is defined as the minimum value of n required to ensure that the errors (either the average error when a prior is present, or the type I and type II errors in the prior-free setting) are below specified thresholds. In this talk, we identify the sample complexity (up to universal multiplicative constants) as a function of p, q, and the desired error thresholds. Our main technical contribution is a novel f-divergence inequality that relates a Jensen-Shannon-type divergence to a Hellinger-type divergence.

Po-Ling Loh

Title: Robust empirical risk minimization via Newton's method

Abstract: We study a variant of Newton's method for empirical risk minimization, where at each iteration of the optimization algorithm, we replace the gradient and Hessian of the objective function by robust estimators taken from existing literature on robust mean estimation for multivariate data. After proving a general theorem about the convergence of successive iterates to a small ball around the population-level minimizer, we study consequences of our theory in generalized linear models, when data are generated from Huber's epsilon-contamination model and/or heavy-tailed distributions. We also propose an algorithm for obtaining robust Newton directions based on the conjugate gradient method, which may be more appropriate for high-dimensional settings, and provide conjectures about the convergence of the resulting algorithm.

Compared to the robust gradient descent algorithm proposed by Prasad et al. (2020), our algorithm enjoys the faster rates of convergence for successive iterates often achieved by second-order algorithms for convex problems, i.e., quadratic convergence in a neighborhood of the optimum, with a stepsize which may be chosen adaptively via backtracking linesearch.

This is a joint work with Eirini Ioannou (Edinburgh) and Muni Sreenivas Pydi (Paris Dauphine - PSL)

Marco Mondelli (Talk 1).

Title: Inference in High Dimensions for Generalized Linear Models: From Spectral Estimators to Approximate Message Passing… And Back

Abstract: In a generalized linear model (GLM), the goal is to estimate a d-dimensional signal x from an n-dimensional observation of the form f(Ax, w), where A is a design matrix and w is a noise vector. Well-known examples of GLMs include linear regression, phase retrieval, 1-bit compressed sensing, and logistic regression. We focus on the high-dimensional setting in which both the number of measurements n and the signal dimension d diverge, with their ratio tending to a fixed constant. Spectral methods provide a popular solution to obtain an initial estimate, and they are also commonly used as a 'warm start' for other algorithms. In particular, the spectral estimator is the principal eigenvector of a data-dependent matrix, whose spectrum exhibits a phase transition.

In the talk, I will start by discussing the emergence of this phase transition and provide precise asymptotics on the high-dimensional performance of spectral methods.
Next, I will combine spectral methods with Approximate Message Passing (AMP) algorithms, thus solving a key problem related to their initialisation. Finally, I will consider two instances of GLMs that incorporate additional structure: (i) a mixed GLM with multiple signals to recover, which offers a flexible solution in settings with unlabelled heterogeneous data, and (ii) a GLM with structured measurement vectors, where the structure is captured by a non-trivial covariance matrix. To study spectral estimators in these challenging cases, the plan is to go back to Approximate Message Passing: I will demonstrate that the AMP framework not only gives Bayes-optimal algorithms, but it also unveils phase transitions in the spectrum of random matrices, thus leading to a precise asymptotic characterisation of spectral estimators.

Marco Mondelli (Talk 2)

Title: Understanding Gradient Descent for Over-parameterised Deep Neural Networks: Insights from Mean-Field Theory and the Neural Tangent Kernel

Abstract: In this second talk, I will depart from the generalized linear model considered earlier to focus on neural networks. Due to the complexity of the parametric model, I will not aim at a precise asymptotic characterization, but rather study the behaviour of gradient descent methods, which constitute the workhorse of deep learning.

Training a neural network is a non-convex problem that exhibits spurious and disconnected local minima. Yet, in practice neural networks with millions of parameters are successfully optimized via gradient descent. I will give some theoretical insights on why this is possible and discuss two approaches to study the behavior of gradient descent. Both approaches crucially exploit the high-dimensionality of the problem — in the same spirit of the previous lecture.

Training a neural network is a non-convex problem that exhibits spurious and disconnected local minima. Yet, in practice neural networks with millions of parameters are successfully optimized via gradient descent. I will give some theoretical insights on why this is possible and discuss two approaches to study the behavior of gradient descent. Both approaches crucially exploit the high-dimensionality of the problem — in the same spirit of the previous lecture.

The first approach takes a mean-field view and it relates the dynamics of stochastic gradient descent (SGD) to a certain Wasserstein gradient flow in probability space. I will show how this idea allows to study the connectivity, convergence and implicit bias of the solutions found by SGD. In particular, I will focus on a recent work proving that, among the many functions that interpolate the data, ReLU networks at convergence implement a simple piecewise linear map of the inputs. The second approach consists in the analysis of the Neural Tangent Kernel. I will show how to bound its smallest eigenvalue in deep networks with minimum over-parameterization, and discuss implications on memorization, optimization, and robustness.

Michèle Wigger

Title: Information-Theoretic Limits of Distributed Detection Systems

Abstract: Distributed detection systems are system-critical components of many modern information processing systems, for example in building maintenance, healthcare, environmental monitoring, etc. In this talk I will present some classic information-theoretic results on distributed hypothesis testing as well as new results on multi-user extensions and on systems with security constraints.