Youth in high dimensions 2022

Abstract book

Ada Altieri (Université Paris Cité)

Alberto Bietti (NYU)

Elisabetta Cornacchia (EPFL)

Nicolas Flammarion (EPFL)

Laura Foini (CNRS)

Mario Geiger (MIT)

Jonathan Kadmon (The Hebrew University)

Michael Lindsey (NYU)

Jihao Long (Princeton University)

Benjamin McKenna (IST Austria)

Cengiz Pehlevan (Harvard)

Noam Razin (Tel Aviv University)

Dominik Schröder (ETH Zürich)

Mark Sellke (Stanford)

Gabriele Sicuro (King's College, London)

Christos Thrampoulidis (University of British Columbia)

Yuting Wei (University of Pennsylvania)

Fanny Yang (ETH Zürich)

Yizhe Zhu (University of California, Irvine)

Ada Altieri (Université Paris Cité)

Title: Dynamical Mean-Field Theory: from glassy systems to ecology and inference

Abstract: Incredible number of metastable states, slow dynamics, and aging are hallmarks of glassiness. Since the seminal works by Sompolinsky and Zippelius in the eighties, the study of mean-field models has been instrumental in revealing these features.

By focusing on systems with soft disordered interactions — from which many relevant models can be recovered as particular cases— I will show how to incorporate the mean-field theory of aging within Dynamical Mean-Field Theory. I will present both the case with only one slow timescale and that with an infinite hierarchy of slower and slower timescales. More interestingly, for the latter, I will discuss how to obtain a dynamical derivation of the stochastic equation that precisely corresponds, for the static counterpart, to the full replica symmetry breaking picture proposed by Parisi. These outcomes extend the realm of out-of-equilibrium mean-field scenarios to all situations where DMFT holds with timely applications in amorphous systems under shear deformations, deep neural networks, ecology and evolution.

Alberto Bietti (NYU)

Title Learning Single-Index Models with Shallow Neural Networks

Abstract: Single-index models are a class of functions given by an unknown univariate "link" function applied to an unknown one-dimensional projection of the input. These models are particularly relevant in high dimension, when the data might present low-dimensional structure that learning algorithms should adapt to. While several statistical aspects of this model, such as the sample complexity of recovering the relevant (one-dimensional) subspace, are well-understood, they rely on tailored algorithms that exploit the specific structure of the target function. In this work, we introduce a natural class of shallow neural networks and study its ability to learn single-index models via gradient descent. More precisely, we consider shallow networks in which biases of the neurons are frozen at random initialization. We show that the corresponding optimization landscape is benign, which in turn leads to generalization guarantees that match the optimal sample complexity of dedicated semi-parametric methods.

Elisabetta Cornacchia (EPFL)

Title: An initial alignment between neural network and target is needed for gradient descent to learn.

Abstract: In this work, we introduce the notion of ``Initial Alignment'' (INAL) between a neural network at initialization and a target function. It is proved that if a network and target function do not have a noticeable INAL, then noisy gradient descent on a fully connected network with normalized i.i.d. initialization will not learn in polynomial time. Thus a certain amount of knowledge about the target (measured by the INAL) is needed in the architecture

design. This also provides an answer to an open problem posed in [Abbe and Sandon, 2020]. The results are based on deriving lower-bounds for descent algorithms on symmetric neural networks without explicit knowledge of the target function beyond its INAL.
Joint work with E. Abbé, J. Hazla, C. Marquis.

Nicolas Flammarion (EPFL)

Title: The role of stochasticity in learning algorithms

Abstract: It has been observed that noise induced by Stochastic Gradient Descent when training neural networks generally enhances generalisation performance in comparison to full-batch training. In this talk, we will try to understand how SGD noise biases the training dynamics towards specific prediction functions for regression tasks. More precisely, we will first show that the dynamics of SGD over diagonal linear networks converges towards a sparser linear estimator than the one retrieved by GD. Then, we will show that adding label noise similarly biases the dynamics towards implicitly solving a Lasso program. Our findings highlight the fact that structured noise can induce better generalisation and they help explain the greater performances of stochastic dynamics over deterministic ones, as observed in practice.

Laura Foini (CNRS)

Title: Annealed averages in spin and matrix models

Abstract: A disordered system is denominated ‘annealed’ when the interactions themselves may evolve and adjust their values to lower the free energy. The opposite (‘quenched’) situation when disorder is fixed, is the one relevant for physical spin-glasses, and has received vastly more attention. Other problems however are more natural in the annealed situation: in our work we discuss examples where annealed averages are interesting, in the context of matrix models. I will discuss how in practice, when system and disorder adapt together, annealed systems develop ‘planted’ solutions spontaneously, as the ones found in the study of inference problems. As an application I will consider the probability distribution of elements of a matrix derived from a rotationally invariant (not necessarily Gaussian) ensemble, a problem that maps into the annealed average of a spin glass model.

L. Foini, J. Kurchan, SciPost Phys. 12, 080 (2022)

Mario Geiger (MIT)

Title: Group Theory for Machine Learning

Abstract: Imposing neural network features to be representations of a group restricts the space of function to the equivariant functions. It also gives a structure to the features which might be useful for interpretability purposes. Finally it is empirically observed that equivariant neural networks have a steeper log-log learning curve.

Jonathan Kadmon (The Hebrew University)

Title Order from chaos: computation and learning in cortical networks

Abstract Neural activity in the living brain is noisy and unreliable. Nevertheless, the brain can produce accurate and reliable behavior. In this talk, I will suggest that chaotic activity is favorable for learning dynamic patterns and show how chaotic cortical circuits can learn and output reliable signals. Using dynamic mean-field theory, I will show how a neural network can suppress chaos in a low-dimensional subspace of its activity. In addition, I will present a plausible learning mechanism for low-dimensional dynamics and argue that the cerebellum—an evolutionary-conserved subcortical region—plays a crucial role in cortical learning.

Michael Lindsey (NYU)

Title: Thermal state sampling for numerical linear algebra

Abstract: How do you estimate the diagonal of an implicitly defined matrix with only O(1) matrix-vector multiplications, or in O(n) time? The best existing randomized approaches (based on the randomized SVD and/or Hutchinson-type estimators) fail when both the rank of the matrix and the off-diagonal contributions grow with n. We introduce a new tool for estimating positive definite matrix diagonals as well as, more generally, `thermal averages' of the form Tr[AX], where X is positive definite. The algorithm is adapted from the minimally entangled typical thermal state (METTS) approach for solving finite-temperature quantum many-body problems at finite temperature. While this method was originally introduced in an exponentially high-dimensional setting where vectors cannot be stored explicitly, it can be simplified and greatly improved in more classical numerical linear algebra settings. We discuss the theory of this approach and highlight applications from diverse fields where this technique can improve the scaling of existing algorithms.

Jihao Long (Princeton University)

Title: Perturbational Complexity by Distribution Mismatch: A Systematic Analysis of Reinforcement Learning in Reproducing Kernel Hilbert Space

Abstract: Most existing theoretical analysis of reinforcement learning (RL) is limited to the tabular setting or linear models due to the difficulty in dealing with function approximation in high dimensional space with an uncertain environment. This talk will offer some fresh insight into this challenge by analyzing reinforcement learning in a general reproducing kernel Hilbert space (RKHS). We consider a family of Markov decision processes of which the reward functions lie in the unit ball of an RKHS and transition probabilities lie in an arbitrary set. We define a quantity called "perturbational complexity by distribution mismatch" to characterize the complexity of the admissible state-action distribution space in response to a perturbation in the RKHS. We show that the perturbational complexity gives both the lower bound of the error of all possible algorithms and the upper bound of two specific algorithms for the RL problem. Hence, the decay of perturbational complexity with respect to the perturbation scale gives an informative difficulty measure of the RL problem. We will provide

some concrete examples and discuss whether the complexity decays fast or not in these examples.

Benjamin McKenna (IST Austria)

Title: Landscape complexity beyond invariance
Abstract: The Kac-Rice formula allows one to study the complexity of high-dimensional Gaussian random functions (meaning asymptotic counts of critical points) via the determinants of large random matrices. To date, it has most frequently been used for highly symmetric models like spherical spin glasses, but we develop techniques for studying models with fewer symmetries, which we illustrate on a simple signal-plus-noise model. Joint work with Gérard Ben Arous and Paul Bourgade.

Cengiz Pehlevan (Harvard)

Title: Deep learning theory at limits
Abstract: A strategy for gaining insight into the complexity of deep learning is to study non-trivial yet tractable limits. I will describe our work on three such limits with a focus on results on representation learning and inductive biases.

Noam Razin (Tel Aviv University)

Title: Generalization in Deep Learning Through the Lens of Implicit Rank Lowering
Abstract:
The mysterious ability of neural networks to generalize is believed to stem from an implicit regularization — a tendency of gradient-based optimization to fit training data with predictors of low "complexity." Despite vast efforts, a satisfying formalization of this intuition is lacking. In this talk I will present a series of works theoretically analyzing the implicit regularization in matrix and tensor factorizations, known to be equivalent to certain linear and non-linear neural networks, respectively. Through dynamical characterizations I will establish an implicit regularization towards low rank (for corresponding notions of rank), different from any type of norm minimization, in contrast to prior beliefs. I will then discuss implications of this finding to both theory (possible explanation for generalization over natural data) and practice (compression of neural network layers, novel regularization schemes). Overall, our results highlight the potential of ranks to explain and improve generalization in deep learning.

Works covered in this talk were done in collaboration with Asaf Maman and Nadav Cohen.

Dominik Schröder (ETH Zürich)

Title: Random matrix resolvent analysis via cumulant expansion

Abstract: We give a general introduction to cumulant expansion based methods of deriving self-consistent equations for the resolvents random matrices. We discuss the case of correlated random matrices leading to the matrix Dyson equation, and the case of entry-wise application of non-linear functions leading to scalar quartic equations.

Mark Sellke (Stanford)

Title: A Universal Law of Robustness via Isoperimetry

Abstract: Classically, data interpolation with a parametrized model class is possible as long as the number of parameters is larger than the number of equations to be satisfied. A puzzling phenomenon in deep learning is that models are trained with many more parameters than what this classical theory would suggest. We propose a theoretical explanation for this phenomenon. We prove that for a broad class of data distributions and model classes, overparametrization is necessary if one wants to interpolate the data smoothly. Namely we show that smooth interpolation requires d times more parameters than

mere interpolation, where d is the ambient data dimension. We prove this universal law of robustness for any smoothly parametrized function class with polynomial size weights, and any covariate distribution verifying isoperimetry. In the case of two-layers neural networks and Gaussian covariates, this law was conjectured in prior work by Bubeck, Li and Nagaraj. We also give an interpretation of our result as an improved generalization bound for model classes consisting of smooth functions.

Gabriele Sicuro (King's College, London)

Title: The planted matching problem and its variations
Abstract: I will discuss the problem of inferring a weighted perfect matching hidden in a weighted graph, under the assumption that the distribution of the weights of the hidden matching edges is different from the one of the weights of the other edges in the graph. Using belief propagation, the maximum-a-posteriori estimator is computed. It is found that the problem exhibits a continuous (and possibly infinite-order) phase transition between a partial recovery phase and a full recovery phase with respect to the signal-to-noise ratio. I will briefly comment on a generalisation of the problem, namely the planted k-factor problem.

Christos Thrampoulidis (University of British Columbia)

Title: Finding Structures in Large Models: Imbalance Trouble

Abstract: What are the unique structural properties of models learned by training deep nets to zero training-error? Is there an implicit bias towards solutions of certain geometry? How does this vary across training instances, architectures, and data? Towards answering these questions, the recently discovered Neural Collapse phenomenon formalizes simple geometric properties of the learned embeddings and of the classifiers of deep nets, which appear to be "cross-situational invariant" across architectures and different balanced classification datasets.

But what happens when classes are imbalanced? Is there a (ideally equally simple) description of the geometry that is invariant across class-imbalanced datasets? By characterizing the global optima of an unconstrained-features model, we formalize a new geometry that remains invariant across different imbalance levels. Importantly, it, too, has a simple description despite the asymmetries imposed by data imbalances on the geometric properties of different classes. Overall, we show that it is possible to extend the scope of the neural-collapse phenomenon to a richer class of geometric structures. We also motivate further investigations into the impact of class imbalances on the implicit bias of first-order methods and into the potential connections between such geometry structures and generalization.

Yuting Wei (University of Pennsylvania)

Title: Minimum L1-norm interpolators: Precise asymptotics and multiple descent

Abstract: An evolving line of machine learning works observe empirical evidence that suggests interpolating estimators --- the ones that achieve zero training error --- may not necessarily be harmful. In this talk, we pursue theoretical understanding for an important type of interpolators: the minimum L1-norm interpolator, which is motivated by the

observation that several learning algorithms favor low L1-norm solutions in the over-parameterized regime. Concretely, we consider the noisy sparse regression model under Gaussian design, focusing on linear sparsity and high-dimensional asymptotics (so that both the number of features and the sparsity level scale proportionally with the sample size).

We observe, and provide rigorous theoretical justification for, a curious multi-descent phenomenon; that is, the generalization risk of the minimum L1-norm interpolator undergoes multiple (and possibly more than two) phases of descent and ascent as one increases the model capacity. This phenomenon stems from the special structure of the minimum L1-norm interpolator as well as the delicate interplay between the over-parameterized ratio and the sparsity, thus unveiling a fundamental distinction in geometry from the minimum L2-norm interpolator. Our finding is built upon an exact characterization of the risk behavior, which is governed by a system of two non-linear equations with two unknowns.

Fanny Yang (ETH Zürich)

Title: Fast rates for noisy interpolation require rethinking the effects of inductive bias

Talk: Modern machine learning has uncovered an interesting observation: large overparameterized models can achieve good generalization performance despite interpolating noisy training data. In this talk, we study high-dimensional linear models and show how interpolators can achieve fast statistical rates when their structural bias is moderate. More concretely, while minimum-l2-norm interpolators cannot recover the signal in high dimensions, minimum-l1-interpolators with strong sparsity bias are much more sensitive to noise. In fact, we show that even though they are asymptotically consistent, minimum-l1-norm interpolators converge with a logarithmic rate much slower than the O(1/n) rate of regularized estimators. In contrast, minimum-lp-norm interpolators with 1<p<2 can trade off these two competing trends to yield polynomial rates close to O(1/n).

Yizhe Zhu (University of California, Irvine)

Title: Non-backtracking spectral clustering in sparse random hypergraphs

Abstract: The stochastic block model has been one of the most fruitful research topics in community detection and clustering. Recently, community detection on hypergraphs has become an important topic in higher-order network analysis. We consider the detection problem in a sparse random tensor model called the hypergraph stochastic block model (HSBM). We prove that a spectral method based on the non-backtracking operator for hypergraphs works with high probability down to the generalized Kesten-Stigum detection threshold conjectured by Angelini et al (2015). We characterize the spectrum of the non-backtracking operator for the sparse random hypergraph and provide an efficient dimension reduction procedure using the Ihara-Bass formula for hypergraphs. As a result, the community detection problem can be reduced to an eigenvector problem of a non-normal matrix constructed from the adjacency matrix and the degree matrix of the hypergraph. Based on joint work with Ludovic Stephan (EPFL).