# Youth in High-Dimensions: Recent Progress in Machine Learning, High-Dimensional Statistics and Inference | (smr 3841)

# ABSTRACTS BOOK

**Silvia Pappalardi: The Eigenstate Thermalization Hypothesis and Free Probability**

The Eigenstate-Thermalization-Hypothesis (ETH) has been established as the general framework to understand quantum statistical mechanics. Only recently has attention been paid to so-called general ETH, which accounts for higher-order correlations among matrix elements. In this talk, I will present the close relation between this perspective on ETH and Free Probability theory, as applied to a thermal ensemble or an energy shell. This mathematical framework allows one to reduce in a straightforward way higher-order correlation functions to a decomposition given by minimal blocks, identified as free cumulants, for which we give an explicit formula. I will illustrate examples on two classes of local non-integrable (chaotic) quantum many-body systems: spin chain Hamiltonians and Floquet brickwork unitary circuits. The results show that the non-trivial frequency dependence of free cumulants encodes the physical properties of local many-body systems and distinguishes them from structureless, rotationally invariant ensembles of random matrices. The present results uncover a direct connection between the Eigenstate Thermalization Hypothesis and the structure of Free Probability, widening considerably the latter’s scope and highlighting its relevance to quantum thermalization.

**Marin Bukov: Introduction to Deep Reinforcement Learning with applications in Quantum Control**

Reinforcement learning (RL) is, alongside supervised and unsupervised learning, one of the three main pillars of modern machine learning. In RL, an artificial intelligence agent interacts with its environment in order to solve a task, by maximizing a reward signal. I will give an intuitive introduction to the basics of reinforcement learning and its mathematical framework (environment, states, actions, rewards, Markov decision processes, etc.). We will then discuss simple RL algorithms, such as policy gradient and Q-learning, recently used to teach AI agents to control quantum systems.

**Pietro Rotondo: Statistical mechanics of deep learning beyond the infinite-width limit**

Decades-long literature testifies to the success of statistical mechanics at clarifying fundamental aspects of deep learning. Yet the ultimate goal remains elusive: we lack a complete theoretical framework to predict practically relevant scores, such as the train and test accuracy, from knowledge of the training data. Huge simplifications arise in the infinite-width limit, where the number of units $N_\ell$ in each hidden layer ($\ell=1,\dots, L$, being $L$ the finite depth of the network) far exceeds the number $P$ of training examples. This idealisation, however, blatantly departs from the reality of deep learning practice, where training sets are larger than the widths of the networks. Here, we show one way to overcome these limitations. The partition function for fully-connected architectures, which encodes information about the trained models, can be evaluated analytically with the toolset of statistical mechanics. The computation holds in the ``thermodynamic limit" where both $N_\ell$ and $P$ are large and their ratio $\alpha_\ell = P/N_\ell$, which vanishes in the infinite-width limit, is now finite and generic. This advance allows us to obtain (i) a closed formula for the generalisation error associated to a regression task in a one-hidden layer network with finite $\alpha_1$; (ii) an approximate expression of the partition function for deep architectures (technically, via an ``effective action" that depends on a finite number of ``order parameters"); (iii) a link between deep neural networks in the proportional asymptotic limit and Student's $t$ processes; (iv) a simple criterion to predict whether finite-width networks (with ReLU activation) achieve better test accuracy than infinite-width ones. As exemplified by these results, our theory provides a starting point to tackle the problem of generalisation in realistic regimes of deep learning.

**Ludwig Hruza: How Free Probability appears in mesoscopic quantum systems**

Probability theory is the basic mathematical framework of statistical mechanics and deals with commuting random variables. In the case where the random variables do not commute – for example large random matrices - there is an analogous theory: Free probability theory. Its philosophy is to replace the concept of “independence” by “freeness”, a concept that includes both, the notion of probabilistic independence and algebraic independence. Curiously, this free probability theory becomes relevant when studying transport in mesoscopic quantum systems. In these systems of intermediate size, particles conserve their ability to (quantum mechanically) interfere over large distances, but the microscopic interaction between particles is complex enough (i.e. non-integrable) to produce an effective noise which leads to diffusive transport.
After a purely mathematical introduction to free probability theory, concentrating on its combinatorial flavour in terms of non-crossing partitions, I will introduce a toy model that captures relevant properties of mesoscopic quantum systems, the Quantum Symmetric Simple Exclusion Process (QSSEP). I will explain why fluctuations of quantum coherences in this model (i.e. off-diagonal entries in the density matrix) can be interpreted as the so-called free cumulants in free probability theory. If time permits, I can discuss how free probability helps to find exact results for the entanglement structure in this model.

References:

Hruza, Ludwig, and Denis Bernard. "Coherent fluctuations in noisy mesoscopic systems, the open quantum ssep, and free probability." *Physical Review X* 13.1 (2023): 011045. (https://arxiv.org/abs/2204.11680)

Bernard, Denis, and Ludwig Hruza. "Exact Entanglement in the Driven Quantum Symmetric Simple Exclusion Process." *arXiv preprint arXiv:2304.10988* (2023). (https://arxiv.org/abs/2304.10988)
Speicher, Roland. "Lecture Notes on" Free Probability Theory"." *arXiv preprint arXiv:1908.08125* (2019). (https://arxiv.org/abs/1908.08125)

**Reza Gheissari: High-dimensional limit theorems for stochastic gradient descent**

Stochastic gradient descent (SGD) is the go-to method for large-scale optimization problems in modern data science. Often, these settings are data constrained, so the sample size and the dimension of the parameter space have to scale together. In this "high-dimensional" scaling, we study pathwise limits of the SGD. Namely, we show limit theorems for trajectories of reasonable summary statistics (finite-dimensional functions of the SGD) as the dimension goes to infinity and the step size simultaneously goes to zero. The limits that arise can be complicated ODE's or SDE's, depending on the initialization, step-size and space rescaling. We present these general results, and then discuss their implications for some concrete tasks including classification of XOR Gaussian mixtures via two layer networks. Based on joint work with Gerard Ben Arous and Aukosh Jagannath.

**Kabir Chandrasekher: Convergence guarantees for iterative algorithms in a class of nonconvex empirical risk minimization problems**

Fitting a model to data typically involves applying an iterative algorithm to minimize an empirical risk.  However, given a particular empirical risk minimization problem, the process of algorithm selection is often performed via either expensive trial-and-error or appeal to (potentially) conservative worst case efficiency estimates and it is unclear how to compare and contrast algorithms in a principled and meaningful manner.
In this talk, we present one potential avenue to obtain principled comparisons between algorithms.  We provide a framework—based on Gaussian comparison inequalities—to characterize the trajectory of an iterative algorithm run with sample-splitting on a set of nonconvex model-fitting problems with Gaussian data.  We use this framework to demonstrate concrete separations in the convergence behavior of several algorithms as well as to reveal some nonstandard convergence phenomena.

**Rishabh Dudeja: Spectral Universality in Regularized Linear Regression**

Spectral universality refers to the empirical observation that asymptotic properties of a high-dimensional stochastic system driven by a structured random matrix are often determined only by the spectrum (or singular values) of the underlying matrix - the singular vectors are irrelevant provided they are sufficiently ``generic''. Consequently, the properties of the underlying system can be accurately predicted by analyzing the system under the mathematically convenient assumption that the singular vectors as uniformly random (or Haar-distributed) orthogonal matrices. This general phenomenon has been observed in numerous contexts, including statistical physics, communication systems, signal processing, statistics, and randomized numerical linear algebra. In this talk, I will describe recent progress toward a mathematical understanding of this universality phenomenon. In the context of penalized linear regression with strongly convex regularizers, I will describe nearly deterministic conditions on the design (or feature) matrix under which this universality phenomenon occurs. I will show that these conditions can be easily verified for highly structured design matrices constructed with limited randomnesses like randomly subsampled Hadamard transforms and signed incoherent tight frames. Due to this universality result, the performance of regularized least squares estimators on many structured sensing matrices with limited randomness can be characterized using the rotationally invariant sensing model with uniformly random (or Haar distributed) singular vectors as an equivalent yet mathematically tractable surrogate. This talk is based on joint work with Subhabrata Sen and Yue M. Lu.

**Spencer Frei: Implicit regularization and benign overfitting for neural networks in high dimensions**

In this talk we go over some of our recent work on understanding optimization and generalization in neural networks trained by gradient descent. We focus on the setting of two-layer neural networks that are trained by the logistic loss for a binary classification task. In the first part of the talk, we show that when the training data is sufficiently high-dimensional, gradient descent has an implicit bias towards low-rank solutions: although the weight matrix is full rank at (random) initialization, gradient descent rapidly reduces the rank to something that is independent of the number of rows and columns of the weight matrix. We show how the choice of the initialization variance has a significant effect on this low-rank implicit bias of gradient descent. In the second part of the talk, we consider a particular distributional setting where we show that the trained network exhibits 'benign overfitting': the network perfectly fits noisy training data and simultaneously generalizes near-optimally. Our results hold when the training data is sufficiently high-dimensional, and we discuss experiments which suggest this assumption is essential for overfitting to be 'benign'. Based on joint work with Peter Bartlett, Niladri Chatterji, Wei Hu, Nati Srebro, and Gal Vardi.

**Yihan Zhang: Discovering spikes in random matrices using approximate message passing**

In modern statistics, random matrices of large dimensions often arise with a low-dimensional planted structure. Such structure may originate from an informative component that has small effective dimensions such as sparsity, rank, etc. It is of statisticians' interest to extract such components from large noisy matrices. To this end, understanding the spectral properties of spiked random matrices arising from various models becomes crucial. In this talk, I will present a novel proof technique using the approximate message passing theory that allows us to systematically characterize some spectral properties such as the location of outlier eigenvalues and the overlap between principal components and unknown parameters. The power of this approach will be demonstrated in the context of generalized linear models with general Gaussian design for which there exists no off-the-shelf random matrix theory results for the matrices of interest.
Based on joint work in preparation with Hong Chang Ji, Marco Mondelli and Ramji Venkataramanan.

**Berfin Simsek: Finite-Width Neural Networks: A Landscape Complexity Analysis**

In this talk, I will present an average-case analysis of finite-width neural networks through permutation symmetry. First, I will give a new scaling law for the critical manifolds of finite-width neural networks derived from counting all partitions due to neuron splitting from an initial set of neurons. Zooming into a line of critical points, I will discuss intriguing transitions from saddles to minima through non-strict saddle points. Considering the invariance of zero neuron addition, we will derive the scaling law of the zero-loss manifold that is exact for the population loss. The competition between these two scaling laws gives a notion of landscape complexity of finitely overparameterized neural networks. At the onset of overparameterization, the complexity explodes and then gradually decreases with further overparameterization, dropping to zero for infinitely wide networks. Our results enable a quantitative understanding of (linear) mode connectivity. Finally, based on our theory, we propose an `Expand-Cluster' algorithm for neuron pruning in practice.

**Emile Mathieu: Geometric Neural Diffusion Processe**

Denoising diffusion models have proven to be a flexible and effective paradigm for generative modelling. Their recent extension to infinite dimensional Euclidean spaces has allowed for the modelling of stochastic processes. However, many problems in the natural sciences incorporate symmetries and involve data living in non-Euclidean spaces. In this work, we extend the framework of diffusion models to incorporate a series of geometric priors in infinite-dimension modelling. We do so by a) constructing a noising process which admits, as limiting distribution, a geometric Gaussian process that transforms under the symmetry group of interest, and b) approximating the score with a neural network that is equivariant w.r.t. this group. We show that with these conditions, the generative functional model admits the same symmetry. We demonstrate scalability and capacity of the model, using a novel Langevin-based conditional sampler, to fit complex scalar and vector fields, with Euclidean and spherical codomain, on synthetic and real-world weather data.

**Enric Boix: The staircase property and the leap complexity**

Which functions f : {+1,-1}^d \to \R can neural networks learn when trained with SGD? In this talk, we will consider functions that depend only on a low-dimensional projection of the input. We will study the dynamics of two-layer neural networks trained by SGD with O(d) samples and will show that a hierarchical property, the “merged-staircase property”, is both necessary and nearly sufficient for learning in this setting. For the more general setting of O(d^c) samples, we will define the "leap complexity" of the function, and argue that it controls whether the neural network will learn. Based on joint work with Emmanuel Abbe and Theodor Misiakiewicz.

**Federica Gerace: Data-centric AI: Play with Datasets to Improve Machine Learning Performances**

An emerging research line in machine learning is focusing on the role played by data in learning problems. Quite recently, the statistical physics community has started working in the direction of extending its methods to include data structure. In this talk, I will provide some insights on how this extension can contribute to analyzing several machine learning frameworks, especially those meant to mitigate the need of new labelled data. Among them, I will particularly focus on Transfer Learning, by showing how an analytically tractable model of source-target dataset pairs can lead to useful guidelines for Machine Learning practitioners.

**Jeanne Trinquier: Machine-learning-assisted Monte Carlo fails at sampling computationally hard problems**

Several strategies have been recently proposed in order to improve Monte Carlo sampling efficiency using machine learning tools. A recently developed line of research proposed to solve the problem in an elegant and universal way, by machine learning proper MCMC moves using autoregressive models. While these methods allow for very efficient sampling for some models, I will show that it is still a challenge to do efficient sampling on a class of problems that are known to be exponentially hard to sample using conventional local Monte Carlo at low enough temperatures . In particular, we studied the antiferromagnetic Potts model on a random graph, which reduces to the coloring of random graphs at zero temperature. We tested several machine-learning-assisted Monte Carlo approaches, and we found that they all fail. Our work thus provides good benchmarks for future proposals for smart sampling algorithms.

**Surbhi Goel: Thinking fast with Transformers - Algorithmic Reasoning via Shortcuts**

In this new era of deep learning, the emergent algorithmic reasoning capabilities of Transformer models have led to significant advancements in natural language processing, program synthesis, and theorem proving. Despite their widespread success, the underlying reasons for their efficacy and the nature of their internal representations remain elusive. In this talk, we take the lens of learning the dynamics of finite-state machines (automata) as the underlying algorithmic reasoning task and shed light on how shallow, non-recurrent Transformer models emulate these recurrent dynamics. By employing tools from circuit complexity and semigroup theory, we characterize "shortcut" solutions that allow a shallow Transformer to precisely replicate $T$ computational steps of an automaton with only $o(T)$ layers. We show that Transformers are efficiently able to represent these “shortcuts” using their parameter-efficient ability to compute sparse functions and averages. Furthermore, through synthetic experiments, we confirm that standard training successfully discovers these shortcuts. We conclude with highlighting the brittleness of these “shortcuts” in out-of-distribution scenarios. *This talk is based on joint work with Bingbin Liu, Jordan T. Ash, Akshay Krishnamurthy, and Cyril Zhang.*

**Francesco Cagnetta: Learning hierarchical compositionality with deep convolutional networks: insights from a Random Hierarchy Model**

While deep learning methods have achieved remarkable success in fields as diverse as language processing and protein structure prediction, understanding their inner workings remains a significant challenge. What are the properties of real data that make deep learning methods so successful and how do these methods learn them? Lacking an answer to these questions, we cannot even estimate the order of magnitude of the sample complexity, that is the number of training examples that a method requires to achieve good performance. In this talk, I will explore our approach to solving this critical challenge, which relies on studying a synthetic classification task---the Random Hierarchy Model---built to mimic the hierarchical and compositional structure of natural data. As the subject of an image (e.g. a dog) consists of features (head, body, limbs), themselves consisting of sub-features (eyes, nose and mouth for the head), each of the classes in our model is represented via a hierarchy of randomly-chosen composition rules, whereby high-level features are represented by a number of semantically equivalent strings of sub-features. Thanks to our specific model choice, we are able to predict the sample complexity in terms of the number of classes, the number of semantically equivalent sub-features and the number of composition rules. Intriguingly, our study reveals that the sample complexity is strictly related to the detectability of the correlations between low-level features of the data and their class. Furthermore, it corresponds to the number of data such that the internal representation of a trained deep network becomes invariant to the exchange of semantically equivalent sub-features. Both the existence of correlations between the low-level features of a datum and its label and the insensitivity of learned representations to aspects of the data irrelevant to the task hold well beyond the context of our synthetic classification task. Therefore, our analysis elucidates a general mechanism of deep learning methods.

**Francesco Camilli: Matrix factorization with neural networks of associative memory**

Matrix factorization is an important and challenging mathematical problem encountered in the context of dictionary learning, recommendation systems and machine learning. The study of its Bayes-optimal limits, namely the insurmountable bounds provided by information theory, presents several obstacles that are still hard to overcome. In this talk, I will abandon Bayes-optimality, in favor of an alternative procedure, called “decimation". Decimation is shown to map matrix factorization into a sequence of neural network models of associative memory, of which the Hopfield model is a celebrated example. Each of these networks turn out to depend on the order parameters of the previous ones, that are in turn linked to their retrieval performances. Although sub-optimal in general, decimation has the benefit of completely analyzable performances. Finally, I will exhibit an “oracle” algorithm based on the ground-state search of a neural network, which shows performances that match the theoretical prediction.

Based on: Camilli, Francesco, and Marc Mézard. "Matrix factorization with neural networks." arXiv preprint arXiv:2212.02105 (2022).