The Abdus Salam International Centre for Theoretical Physics

Workshop on Learning and Inference from Structured Data: Universality, Correlations and Beyond (3 - 7 July 2023)

SPEAKERS' Abstracts

Soledad Villar

Title: Approximately equivariant graph networks

Abstract: Graph neural networks (GNNs) are commonly described as being permutation equivariant with respect to node relabeling in the graph. This symmetry of GNNs is often compared to the translation equivariance symmetry of Euclidean convolution neural networks (CNNs). However, these two symmetries are fundamentally different: The translation equivariance of CNNs corresponds to symmetries of the fixed domain acting on the image signal
(sometimes known as active symmetries), whereas in GNNs any permutation acts on both the graph signals and the graph domain (sometimes described as passive symmetries). In this work, we focus on the active symmetries of GNNs, by considering a learning setting where signals are supported on a fixed graph. In this case, the natural symmetries of GNNs are the automorphisms of the graph. Since real-world graphs tend to be asymmetric, we relax the notion of symmetries by formalizing approximate symmetries via graph coarsening.
We present a bias-variance formula that quantifies the tradeoff between the loss in expressivity and the gain in the regularity of the learned estimator, depending on the chosen symmetry group. To validate our approach, we conduct extensive experiments on human pose estimation and traffic flow prediction with different choices of symmetries. We show theoretically and empirically that the best generalization performance can be achieved by choosing a suitably larger group than the graph automorphism group, but smaller than the full permutation group.

Takashi Takahashi

Title: Exploring bagging with structured data: Insights from precise asymptotics

Abstract:
Bagging is a standard method in machine learning. The basic idea of bagging is to average the learning results from pseudo data sets made by resampling or subsampling the acquired data. Past literature shows that, in the context of learning generalized linear models (GLMs)

with a ridge(-less) regularization, bagging is equivalent to introducing implicit l2 regularization from theoretical and experimental perspectives.

To further explore the role of bagging in GLMs, we consider bagging in the context of learning from structured data, specifically sparse linear regression, and logistic regression from label-imbalanced data. Under these settings, we precisely characterize the performance of bagging when the input dimension and the data size diverge at the same rate. As a result, we show that bagging might provide benefits beyond simple l2 regularization when the data exhibits structural characteristics.

Justin Ko

Title: Learning Noisy Rank-One matrices: Bayes, Non-Bayes, and Large Deviations

Abstract: We introduce a universality result that reduces the mutual information for inference problems in the mismatched setting to the computation of a modified SK free energy. In the Bayesian optimal setting, the modified SK free energy is identical to the one from low rank matrix factorization. We prove an almost sure large deviations principle for the overlaps between the truth and estimators in both the Bayesian optimal and mismatched setting. As a consequence, we recover the limit of the mutual information in mismatched inference problems. This is upcoming work with Alice Guionnet, Florent Krzakala, and Lenka Zdeborova.

Manuel Saenz

Title: How much does it cost to forget noise structure in low-rank matrix estimation?

Abstract: In this talk, we will delve into the problem of estimating a rank-1 signal that is corrupted by structured rotationally invariant noise. We specifically address the question of how well inference algorithms perform when the statistics of the noise is not known, and therefore, Gaussian noise is assumed. While the matched Bayes-optimal setting with unstructured noise is well understood, the analysis of mismatched problems is relatively limited. Our aim will be to shed some light on the impact of this strong mismatch between the inference model used and the data generating mechanism. Our primary main contributions involve the rigorous analysis of the corresponding Bayes estimator and approximate message passing (AMP) algorithm, both of which mistakenly assume the presence of Gaussian noise. By providing precise asymptotic characterizations for these estimators, we uncover a diverse and surprising phenomenology.

Matteo Marsili

Title: Data that “make sense”

A definition of learning as "making sense of data that make sense" requires an absolute quantitative notion of relevance, that is independent of what is to be learned and how. We have recently proposed [1] one such notion, based on information theoretic arguments, which applies both to data and representations, thus allowing one to define maximally informative samples and optimal learning machines in a context independent manner. This approach to learning clarifies the relation between "making sense" and criticality and that Gaussian data is necessarily structureless, it provides new approaches to high-dimensional inference in the under-sampling regime and guiding principles for the design of machine learning architectures [2].

[1] M Marsili, Y Roudi, Quantifying relevance in learning and inference. Physics Reports 963, 1-43 (2022)
[2] R Xie, M Marsili, Occam learning. arXiv preprint arXiv:2210.13179 (2023)

Federica Gerace

Title: Gaussian Universality of Perceptrons with Random Labels

Abstract: While classical in many theoretical settings — and in particular in statistical physics-inspired works— the assumption of Gaussian i.i.d. input data is often perceived as a strong limitation in the context of statistics and machine learning. In this study, we redeem this line of work in the case of generalized linear classification, a.k.a. the perceptron model, with random labels. We argue that there is a large universality class of high-dimensional input data for which we obtain the same minimum training loss as for Gaussian data with corresponding data covariance. In the limit of vanishing regularization, we further demonstrate that the training loss is independent of the data covariance. On the theoretical side, we prove this universality for an arbitrary mixture of homogeneous Gaussian clouds. Empirically, we show that the universality holds also for a broad range of real datasets.

Bruno Loureiro

Title: Universality and feature learning in two-layer neural networks

Universality is the theorist best friend, justifying the study of mathematically tractable models beyond their original and limited scope. In the context of machine learning, recent Gaussian universality results have provided useful insight in the properties of neural networks in the lazy regime, for instance illustrating how overparametrisation is not necessarily at odds with generalisation. However, it is also limiting, as it is known that in this regime overparametrised neural networks can learn as well as a polynomial kernel in the degree of the sample complexity. In this talk, I will discuss the recent progress in universality results and beyond, when feature learning kicks in and neural networks (might) outperform kernels.

Sebastian Goldt

Title: The Gaussian world is not enough - how data shapes neural network representations

What do neural networks learn from their data ? We discuss this question in two learning paradigms: supervised classification with feed-forward networks, and masked language modelling with transformers. First, we give analytical and experimental evidence for a "distributional simplicity bias", whereby neural networks learn increasingly complex distributions of their inputs. We then show that neural networks learn from the higher-order cumulants (HOCs) more efficiently than lazy methods, and show how HOCs shape the learnt features. We finally characterise the distributions that are learnt by single- and multi-layer transformers, and discuss implications for learning dynamics and transformer design.

Jean Barbier

Title: Fundamental limits in structured PCA, and how to reach them

Abstract: I will present recent results concerning the Bayesian estimation of low-rank matrices corrupted by structured noise, namely rotational invariant noise with generic spectrum. Using the replica method we derive the optimal performance limit. This is possible by exploiting the low-rank structure of the matrix signal implying that we can reduce the model to an effective quadratic model of the Ising type. Secondly, we show that the Approximate Message Passing (AMP) algorithm currently proposed in the literature for Bayesian estimation is sub-optimal. Exploiting the theory of Adaptive Thouless-Anderson-Palmer equations by Opper et al. we explain the reason for this sub-optimality and as a consequence we deduce an optimal Bayesian AMP algorithm with a rigorous state evolution matching the replica prediction.

Santiago Acevedo

Title: Intrinsic dimension estimation in spin systems.

Abstract: We will start by introducing the recently developed intrinsic dimension estimator for discrete spaces[1], and presenting its limitations to work on Ising spin systems. Then, we will show how to modify the algorithm to overcome these limitations, and give examples of performance on thermal phase transitions. Finally, we will discuss possible relationships with the ID in Ising spin systems and the Shannon entropy of the underlying probability distribution.

[1] Macocco et al. Intrinsic Dimension Estimation for Discrete Metrics, Physical Review Letters. (2023)

Manfred Opper

Title: Replica method with approximate inference

The replica method has been used over many years to compute a variety of exact learning curves
in the so-called teacher student scenario for single layer neural network architectures in the thermodynamic limit. Such computations are
made possible by idealised assumptions on simple, often i.i.d. Gaussian distributed input data.

In recent years, there has been an increasing interest to extend replica results to learning with real world data, e.g. by approximating input distributions of network activation functions
using second order statistics determined by empirical data.

In this talk (based on earlier work [1]) I will discuss an alternative framework for making the replica method applicable to learning on real data. Approximating the distribution of the replicated and averaged system using approximate inference methods (such as the variational approach), and by assuming
replica symmetry, one can obtain approximations for effective marginal densities. This approach could possibly deal with more general assumptions on data generation, avoiding the need for specifying explicit 'teacher' networks.

[1] Dörthe Malzahn and Manfred Opper J. Stat. Mech. (2005) P11001

Yoshiyuki Kabashima

Title: Compressed sensing based on diffusion models

Abstract: Compressed Sensing (CS) is a framework that takes advantage of the expected "sparsity" of many natural signals, allowing them to be recovered from fewer measurements. Although the last two decades witnessed many successful applications of CS, it does not provide the optimal sensing scheme. Theoretically, the optimal performance is achieved by Bayes' theorem exploiting the "correct" prior, which has been considered difficult to realize. However, the recent advance in research on generative models makes it possible to numerically construct "reasonable" priors for natural signals. We talk about our recent attempt to improve the performance of CS utilizing diffusion-based generative models as priors.
This is a joint work with Xiangming Meng (Zhejiang University).
References:
X Meng and Y. Kabashima,
arXiv:2302.00919; arXiv:2211.13006 ; arXiv:2211.12343

Marco Mondelli

Title: 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, in the spirit of the workshop title, 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.

Based on a series of joint works with Hong Chang Ji, Andrea Montanari, Ramji Venkataramanan, and Yihan Zhang.

Zhou Fan

Title: On orthogonally-invariant spin glasses and linear models with invariant designs

Abstract: I will discuss some recent progress towards the rigorous analysis of the asymptotic mutual information and TAP mean-field equations in linear models with rotationally-invariant designs, and related analyses of an orthogonally-invariant analogue of the SK model with external field. Our results establish the validity of the replica-symmetric predictions for these quantities at sufficiently high temperatures, using an adaptation of a conditional second-moment argument of Bolthausen, where we condition on the filtrations of iterative algorithms that solve the mean-field equations in such models. This is joint work with Yufan Li, Subhabrata Sen, and Yihong Wu.

Yue M. Lu

Title: Equivalence Principles for Nonlinear Random Matrices: A Closer Look

Abstract: In recent years, several novel random matrix ensembles have emerged in the fields of machine learning and signal processing. The spectral properties of these matrices have been shown in numerous studies to be instrumental in addressing critical issues such as the training and generalization performance of neural networks, and the fundamental limits of high-dimensional signal recovery. Consequently, there is an increasing interest in accurately understanding the spectral and other asymptotic properties of these matrices. Differing from their classical counterparts, these new random matrices often bear a greater degree of structure, a result of nonlinear transformations. This combination of structure and nonlinearity presents substantial technical challenges when attempting to apply existing tools from random matrix theory to these new ensembles.

In this presentation, I will discuss several closely related equivalence principles that establish an asymptotic equivalence between a variety of nonlinear random matrices and certain linear random matrix ensembles, which are comparatively easier to analyze. Furthermore, I will demonstrate how these equivalence principles can be employed to characterize the performance of kernel methods and random feature models across different scaling regimes.

Joint work with Hong Hu and Horng-Tzer Yau

Rishabh Dudeja

Title: Spectral Universality in High-Dimensional Statistics

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.

Pragya Sur

Title: Debiasing regularized linear estimators with ``spectrum-aware adjustments"

Abstract: Debiasing methodologies have emerged as a solid toolbox for producing inference in high-dimensional problems. Since its original introduction, the methodology witnessed a major upheaval with the introduction of debiasing with degrees-of-freedom adjustments that arises from Onsager correction terms in high dimensions. In this talk, we study such degrees-of-freedom corrected debiasing formula for rotationally invariant designs rigorously, building upon the statistical mechanics insights from Takahashi and Kabashima (2018). We name this class of corrections ``spectrum-aware adjustments" to capture their dependence on spectral properties of the design. We demonstrate the utility of such formulae with regard to statistical problems in high-dimensional linear regression such as hypothesis testing, signal-to-noise ratio estimation, etc. Further, we observe the superiority of such corrections over previous Gaussian-based formulae in the context of challenging scenarios where one might encounter dependent observations, multivariate-t-distributions, and noisy low-rank matrices. Finally, this approach integrates seamlessly with the principal components regression (PCR) methodology yielding a new perspective on PCR. This is based on joint work with Yufan Li.

Francesco Camilli

Title: Fundamental limits of shallow neural networks with small training sets

Abstract: In my presentation I shall provide an information-theoretical analysis of a two-layer neural network trained on a relatively small dataset compared to the network size. The dataset is generated by a teacher network with the same architecture. The main finding of this study is the asymptotic equivalence of the two-layer neural network with a Generalized Linear Model, where the Mutual information between the weights and the training set is known. This result in turn yields the Bayes-optimal generalization error, which serves as a lower bound for any neural network with the mentioned architecture. The proof relies on rigorous Mathematical Physics tools used in the study of spin glasses, such as the interpolation scheme, and it is guided by recently conjectured Gaussian Equivalence Principles. With respect to the existing literature, which is either non-rigorous, or restricted to the case of the learning of the readout weights only, the proof addresses the learning of all the network parameters. While our techniques are primarily applicable to the regime of small datasets, it offers the advantage of being self-contained, simple, and it constitutes an independent proof of a Gaussian Equivalence Principle.