

Bologna–SISSA–ICTP mini-workshop on ML and beyond

Programme · 27–28 October 2026 · Euler–Lagrange room, ICTP

Tuesday 27 October

09:15	<i>Welcome</i>	
09:30	Shu Zhou SISSA	<i>Mesoscopic learning dynamics of extensive-width neural networks</i>
10:00	Leonardo Ferreira ICTP & Federal University of Rio Grande do Sul	<i>PCA and fundamental thresholds of an inhomogeneous spiked Wigner model with random variance profile</i>
10:30	<i>Coffee break</i>	
11:00	Giovanni Lucarelli SISSA	<i>Encoding of mathematics in large language models</i>
11:30	Emanuele Mingione University of Bologna	<i>Variational bounds for perceptron learning from structured data</i>
12:00	Eleonora Bergamin SISSA	<i>How to define generalisation for a generative model?</i>
12:30	<i>Lunch</i>	
14:00	Gabriele Sicuro University of Bologna	<i>Atlases of matching perturbations</i>
14:30	Carlo Caputo SISSA	<i>Symmetry breaking in the gradient-flow dynamics of wide two-layer networks near interpolation</i>
15:00	<i>Coffee break</i>	
15:30	<i>Open discussion</i>	
17:00	<i>End of day</i>	

Wednesday 28 October

10:00	Mauro Pastore ICTP	<i>Feature norm optimization and geometry as driving mechanism of neural scaling law</i>
10:30	Santiago Acevedo SISSA	<i>Renormalizing language representations: scale invariance and universality in LLMs</i>
11:00	<i>Coffee break</i>	
11:30	<i>Open discussion</i>	
12:30	<i>Lunch</i>	
14:00	Pierluigi Contucci ICTP Colloquium	<i>The language of natural numbers: statistical properties and neural learning of arithmetic structures</i>
15:00	<i>End of workshop</i>	

Abstracts

Mesoscopic learning dynamics of extensive-width neural networks

Shu Zhou (SISSA) · Tue 27, 09:30

Understanding how stochastic gradient descent learns internal representations in wide neural networks trained on rich target functions is a central problem in machine learning. We study this question in two-layer neural networks learning target functions represented by networks of the same architecture, with arbitrary nonlinearity, and with network width and input dimension growing proportionally. In this regime, standard approaches for describing the learning dynamics require tracking a diverging number of overlaps, which measure correlations between learned and target features, leaving no obvious description in terms of a few order parameters. Here, we develop a mesoscopic description of learning by characterizing the distribution of overlaps. In the early stages of learning, we find that a Gaussian description of overlap statistics reduces the learning dynamics to a few collective variables that describe the acquisition of the target function's linear and quadratic components. We then estimate the timescale of specialization, when Gaussian closure breaks down and individual learned features begin to strongly correlate with specific target features. We finally describe how the activation function determines the pattern of specialization and characterize the subsequent refinement of learned features. Together, these results clarify how collective learning gives way to individual feature discovery and provide a framework for understanding when learning in large neural networks admits a simple statistical description.

PCA and fundamental thresholds of an inhomogeneous spiked Wigner model with random variance profile

Leonardo Ferreira (ICTP & Federal University of Rio Grande do Sul) · Tue 27, 10:00

The spiked Wigner ensemble is a prototypical model for high-dimensional inference. Its simplicity, combined with the powerful insights it provides, have raised it to a paradigm of low-rank signal recovery in the presence of homogeneous noise. However, recent works have introduced structured versions of the spiked Wigner model in which the standard assumptions of independence and identical distribution are relaxed. In this work, we study the recovery thresholds of an inhomogeneous rank-one spiked Wigner model in which the variance of each entry of the noise matrix is itself a random variable, described by a distribution considered as input. A distinguishing feature of our model, compared with previous approaches in random matrix theory and inhomogeneous signal recovery, are the strong correlations arising from the rank-one structure of our variance matrix. First, we characterize the PCA threshold, defined as the BBP line that separates the gapped phase, where the signal is detectable by means of the spectrum of the observation matrix, from the gapless phase. We derive exact equations for the spectral edges, the outlier eigenvalue, and the distribution of the components of the outlier eigenvector. Then, we characterize the fundamental threshold of the model, determining the regime where it is information-theoretically possible to recover the signal component. We present numerical results for a truncated power-law variance profile, and discuss how noise intensity and heterogeneity affect the recovery thresholds.

Encoding of mathematics in large language models

Giovanni Lucarelli (SISSA) · Tue 27, 11:00

Despite substantial progress in mathematical reasoning, the internal representations underlying LLM performance remain poorly understood. This work investigates the geometry of hidden representations of algebraic expressions in mathematics-specialized LLMs. We examine how operator structure and algebraic equivalence are reflected across model layers. Using neighborhood-based representational similarity, we find that expressions sharing the same operator structure exhibit high similarity across layers, whereas algebraically equivalent expressions with different operator structures show the strongest similarity in intermediate layers. Methodologically, we also show that low point-wise cosine similarity can mask substantial geometric correspondence, which can be recovered through orthogonal Procrustes alignment. This highlights the importance of accounting for global orthogonal reorientations in representational analyses.

Variational bounds for perceptron learning from structured data

Emanuele Mingione (University of Bologna) · Tue 27, 11:30

We consider the Gibbs measure associated to a continuous-spin perceptron trained on a Gaussian mixture for a large class of concave utilities and log-concave, separable prior measures on the spins. By combining the adaptive interpolation method with log-concavity properties of the prior, we obtain lower and upper bounds for the limiting quenched pressure. The bounds are given in terms of a min-max, finite-dimensional, variational problem on the same functional and differ only by the min-max order. Whenever the two bounds match, one can derive the ground-state energy, training loss, and generalization error for the model.

How to define generalisation for a generative model?

Eleonora Bergamin (SISSA) · Tue 27, 12:00

According to an influential recent proposal, an image diffusion model is said to generalise if two models trained independently on disjoint subsets of a dataset generate nearly identical images given the same latent noise. Extending this idea to autoregressive transformers is challenging: unlike diffusion models, transformers lack a shared latent variable that naturally couples generation, and discrete tokens lack an obvious notion of overlap. Here, we introduce the survival probability — the probability that two transformers trained on disjoint subsets of the training data generate the same sequence under maximal coupling — and use it to empirically demonstrate a transition from verbatim memorisation to generalisation in transformers as training set size increases. We further show that independently trained transformers develop a shared predictive subspace, defined via the leading principal components of next-token probabilities, when trained on data sets that are substantially smaller than those required for convergence. We then introduce a simple auto-regressive generative model, the spiked Markov chain, and develop an asymptotically sharp characterisation of convergence and subspace recovery for this model. Our framework provides new tools for assessing generalisation in auto-regressive generative models and motivates investigating how transformer inductive biases shape both the emergence of shared predictive structure and the agreement between generated sequences.

Atlases of matching perturbations

Gabriele Sicuro (University of Bologna) · Tue 27, 14:00

I will present some results on the set of optimal matchings that can be obtained on so-called generically-weighted hypomatchable graphs. Such graphs allow for a perfect matching modulo the removal of a node at a time, which plays the role of a local "impurity/perturbation". The results might shed some light on the statistical properties of the random Euclidean assignment problem on 2d manifolds.

Symmetry breaking in the gradient-flow dynamics of wide two-layer networks near interpolation

Carlo Caputo (SISSA) · Tue 27, 14:30

We study the gradient-flow dynamics of a two-layer student network trained on a fixed dataset generated by a two-layer teacher of possibly different width, in the regime where the number of samples grows with the number of parameters. Building on recent progress by Montanari and Urbani (2025), who analysed gradient flow for a single-index teacher with a symmetric dynamical mean-field theory, we show that for a multi-index teacher this symmetric solution is unstable. On a typical dataset the hidden units break their permutation symmetry and specialise to different teacher directions after a time that grows logarithmically with the input dimension, so an ansatz that breaks the symmetry is needed. We build a sequence of refined ansatzes that break the symmetry between units already at initialisation, while keeping the number of order parameters independent of the width. The resulting theory agrees with gradient descent for generic activations and captures the phenomenology of the different regimes. In the mean-field scaling all the hidden units organise into blocks aligned with different teacher directions. In the standard scaling only a fraction of them is used to learn the features. With inhomogeneous teacher readouts, the units first align with the most important feature and then switch to the less important ones, in a cascade. We then discuss the regimes in which the theory is expected to break down. Finally, we discuss future directions: using the theory to study the escape time from the unspecialised plateau as the number of features to learn grows, the overfitting time scale, and the long-time dynamics.

Feature norm optimization and geometry as driving mechanism of neural scaling law

Mauro Pastore (ICTP) · Wed 28, 10:00

Neural scaling laws describe how the performance of deep learning models follows power-law trends with model size, training data, and compute. Their origin is often attributed to hierarchical or power-law structure in the data or target function. Here, we consider a minimal setting in which no such structure is present, yet scaling laws emerge from the learning problem itself. We study a two-layer teacher–student model in the data-rich, strongly undercomplete regime, where the teacher is much wider than the student. We explain this behavior analytically by separating the amplitude of student features from the geometry of their directions. We show that optimizing the amplitude is sufficient to produce an inverse-width scaling law, while low-order statistics of the learned directions determine its prefactor. This distinction explains how scaling can arise even with random directions, and how learning improves performance through a more favorable representation geometry.

Renormalizing language representations: scale invariance and universality in LLMs

Santiago Acevedo (SISSA) · Wed 28, 10:30

It was recently observed that language representations of LLMs processing sets of translated sentences develop similar neighborhood structures, namely, they become "locally aligned". However, we still lack an understanding of what information is shared by representations in semantic correspondence. We propose the Renormalization Group from Statistical Mechanics as a framework for modeling the information shared between translations. Consistent with this picture, we show that LLM neural activity shows signatures of scale invariance along the token axis only at the central-depth layers, and that cross-lingual representational convergence is strongly concentrated in the long-wavelength modes of LLM neural activity along the text axis.

The language of natural numbers: statistical properties and neural learning of arithmetic structures

Pierluigi Contucci (ICTP Colloquium) · Wed 28, 14:00

The relationship between artificial intelligence and science is increasingly bidirectional: AI offers new tools for discovering and exploring scientific structures, while science provides controlled domains in which to investigate the capabilities and limitations of AI. We examine this interplay in the fundamental setting of arithmetic. Iterated prime factorization turns the integers into a deterministic sequence of words whose vocabulary and admissible combinations are governed by arithmetic. We establish asymptotic laws for vocabulary growth and word frequencies, examine statistical dependencies, and show that a transformer trained solely on such text captures structural regularities, including sensitivity to arithmetically forbidden sequences. This resulting language supports both directions of inquiry: neural models make arithmetic structure accessible through learning, while arithmetic supplies precise constraints against which to examine what these models learn and how they represent it.