The Abdus Salam International Centre for Theoretical Physics

United Nations Educational, Scientific and Cultural Organization

IAEA International Atomic Energy Agency

6th Youth in High-Dimensions: Recent Progress in Machine Learning, High-Dimensional Statistics and Inference & How creative is Generative AI? Perspectives from Science and Philosophy | (SMR 4084)

03 Jul 2025 - 10 Jul 2025
ICTP, Trieste, Italy

T01 - KAMB Mason

An Analytic Theory of Creativity in Convolutional Diffusion Models

T02 - KOVAČEVIĆ Filip

Spectral Estimators for Multi-Index Models: Precise Asymptotics and Optimal Weak Recovery

T03 - MASCARETTI Andrea

What's in a caption? Finding semantics with the information imbalance.

T04 - MEDVEDEV Marko

Weak-to-Strong Generalization Even in Random Feature Networks, Provably.

T05 - SAEED Basil

Local minima of the empirical risk in high-dimensions: general theorems and convex examples

T06 - TIBERI Lorenzo

Dissecting the Interplay of Attention Paths in a Statistical Mechanics Theory of Transformers

T07 - URFIN Raphaël

Memorization and Generalization in Generative Diffusion

T08 - VASUDEVA Bhavya

The Rich and the Simple: On the Implicit Bias of Adam and SGD

T09 - ZHANG Yedi

Training Dynamics of In-Context Learning in Linear Attention

T01

An Analytic Theory of Creativity in Convolutional Diffusion Models

Mason Kamb [1] **, Surya Ganguli**[1]

[1]*Department of Applied Physics, Stanford University*

Score-matching diffusion models can generate highly original images that lie far from their training data. However, optimal score-matching theory suggests that these models should only be able to produce memorized training examples. To reconcile this theory-experiment gap, we identify two simple inductive biases, locality and equivariance, that: (1) induce a form of combinatorial creativity by preventing optimal score-matching; (2) result in fully analytic, completely mechanistically interpretable, local score (LS) and equivariant local score (ELS) machines that, (3) after calibrating a single time-dependent hyperparameter can quantitatively predict the outputs of trained convolution only diffusion models (like ResNets and UNets) with high accuracy (median r^2 of 0.95, 0.94, 0.94, 0.96 for our top model on CIFAR10, FashionMNIST, MNIST, and CelebA). Our model reveals a *locally consistent patch mosaic* mechanism of creativity, in which diffusion models create exponentially many novel images by mixing and matching different local training set patches at different scales and image locations. Our theory also partially predicts the outputs of pre-trained self-attention enabled UNets (median r^2 ~ 0.77 on CIFAR10), revealing an intriguing role for attention in carving out semantic coherence from local patch mosaics.

T02

Spectral Estimators for Multi-Index Models: Precise Asymptotics and Optimal Weak Recovery

Filip Kovačević[1], **Yihan Zhang**[2], **and Marco Mondelli**[1]

[1]*Institute of Science and Technology Austria*
[2]*University of Bristol*

Multi-index models provide a popular framework to investigate the learnability of functions with low-dimensional structure and, also due to their connections with neural networks, they have been object of recent intensive study. In this paper, we focus on recovering the subspace spanned by the signals via spectral estimators – a family of methods routinely used in practice, often as a warm-start for iterative algorithms. Our main technical contribution is a precise asymptotic characterization of the performance of spectral methods, when sample size and input dimension grow proportionally and the dimension p of the space to recover is fixed. Specifically, we locate the top-p eigenvalues of the spectral matrix and establish the overlaps between the corresponding eigenvectors (which give the spectral estimators) and a basis of the signal subspace. Our analysis unveils a phase transition phenomenon in which, as the sample complexity grows, eigenvalues escape from the bulk of the spectrum and, when that happens, eigenvectors recover directions of the desired subspace. The precise characterization we put forward enables the optimization of the data preprocessing, thus allowing to identify the spectral estimator that requires the minimal sample size for weak recovery.

T03

What's in a caption? Finding semantics with the information imbalance.

Andrea Mascaretti[1], **Santiago Acevedo**[1], **Marco Baroni**[2], **Alessandro Laio**[1], **Matéo Mahaut**[2], **and Riccardo Rende**[1]

[1]*(Presenting author underlined) Scuola Internazionale Studi Superiori Avanzati*
[2]*Universitat Pompeu Fabra*

Size is the driving factor behind the success of transformers. Bigger models produce higher-dimensional representations, encode more information, and achieve better performance; examples of this trend are large language models (LLMs) and visual transformers (ViTs). Size also drives convergence: [1] conjecture that representations are effective because they mimic reality. Transformers produce representations that (i) are very high dimensional, and (ii) possess highly-nonlinear relations across layers. Comparing the hidden representations from different transformers is a valuable tool to localise and contrast these semantic representations; and require methods that (i) scale with size, (ii) capture local behaviours, and (iii) allow to quantify asymmetries between different models. Linear methods are fast, but do not capture local behaviours; local methods, that count the number of common neighbours, are symmetric, and do not allow to quantify the partial ordering induced by the scale effects of transformers. We propose a new method based on the information imbalance: a local and asymmetric measure of topological similarity. We employ the method to study the hidden representation of DeepSeekV3, a state-of-the-art LLM, and study the relation between hidden representations and semantics in (i) a text-text scenario, in which we contrast sentences in different languages; (ii) a text-image scenario, in which we compare images and their captions. The method allows to investigate the data employing an arbitrary number of tokens and bigger sample sizes. We find that (i) alignment is maximal at semantic layers, regardless of the specific transformer or domain; (ii) the auto-correlation of tokens is higher in the semantic regions of transformers; (iii) the image-text transformer have asymmetric information imbalances. The findings support the idea of a representational convergence; but also highlights the importance of building tools that account for the size of the representations and the uneven distribution of information across layers and tokens.

[1] Huh, Minyoung, et al. "The platonic representation hypothesis." arXiv preprint arXiv:2405.07987 (2024).

T04

Weak-to-Strong Generalization Even in Random Feature Networks, Provably.

Marko Medvedev[1], **Kaifeng Lyu**[2], **Dingli Yu**[3], **Sanjeev Arora**[4], **Zhiyuan Li**[5], **Nathan Srebro**[5]

[1]*The Unviersity of Chicago,* [2]*Simons Institute, University of California,* [3]*Microsoft Research,* [4]*Princeton University,* [5]*TTIC*

Weak-to-Strong Generalization (Burns et al., 2023) is the phenomenon whereby a strong student, say GPT-4, learns a task from a weak teacher, say GPT-2, and ends up significantly outperforming the teacher. We show that this phenomenon does not require a strong learner like GPT-4. We consider student and teacher that are random feature models, described by two-layer networks with a random and fixed bottom layer and trained top layer. A 'weak' teacher, with a small number of units (i.e. random features), is trained on the population, and a 'strong' student, with a much larger number of units (i.e. random features), is trained only on labels generated by the weak teacher. We demonstrate, prove and understand, how the student can outperform the teacher, even though trained only on data labeled by the teacher. We also explain how such weak-to-strong generalization is enabled by early stopping. Importantly, we also show the quantitative limits of weak-to-strong generalization in this model.

This work will appear as a poster in ICML 2025.

[1] Burns, C., Izmailov, P., Kirchner, J. H., Baker, B., Gao, L., Aschenbrenner, L., Chen, Y., Ecoffet, A., Joglekar, M., Leike, J., Sutskever, I., and Wu, J. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=ghNRg2mEgN.

T05

Local minima of the empirical risk in high-dimensions: general theorems and convex examples

Kiana Asgari[1], **Andrea Montanari**[1], **and Basil Saeed**[1]

[1]*Stanford University*

We consider a general model for high-dimensional empirical risk minimization:

$$\hat{\boldsymbol{\Theta}}_n \in \arg\min_{\boldsymbol{\Theta}\in\mathbb{R}^{d\times k}} \hat{R}_n(\boldsymbol{\Theta}), \qquad \hat{R}_n(\boldsymbol{\Theta}) := \frac{1}{n}\sum_{i=1}^{n} L(\boldsymbol{f}(\boldsymbol{\Theta}^{\mathsf{T}}\boldsymbol{x}_i), \boldsymbol{y}_i) + r(\boldsymbol{\Theta})$$

where the data $\boldsymbol{x}_i \in \mathbb{R}^d$ are Gaussian vectors, $\boldsymbol{y}_i \in \mathbb{R}^q$ with $\mathbb{P}(\boldsymbol{y}_i \in \mathcal{B}|\boldsymbol{x}_i) = g(\mathcal{B}|\boldsymbol{\Theta}_0^{\mathsf{T}}\boldsymbol{x}_i)$ for some $\boldsymbol{\Theta}_0 \in \mathbb{R}^{d\times k_0}$, the regularizer is separable, the model is parameterized by $\boldsymbol{\Theta} \in \mathbb{R}^{d\times k}$, and the loss depends on the data via the projection $\boldsymbol{\Theta}^{\mathsf{T}}\boldsymbol{x}_i$ potentially in a non-convex manner. This setting covers, as special cases, classical statistics methods (e.g. multinomial regression and other generalized linear models), but also two-layer fully connected neural networks with k hidden neurons.

We use the Kac–Rice formula from Gaussian process theory to derive a bound on the expected number of local minima of this empirical risk, under the proportional asymptotics in which $n, d \to \infty$, with $n \asymp d$. Namely, let $\hat{\mu}_{\sqrt{d}[\boldsymbol{\Theta},\boldsymbol{\Theta}_0]} \in \mathscr{P}(\mathbb{R}^{k+k_0})$, $\hat{\nu}_{[\boldsymbol{X}\boldsymbol{\Theta},\boldsymbol{y}]} \in \mathscr{P}(\mathbb{R}^{k+q})$ denote the empirical distributions of the rows of $\sqrt{d}[\boldsymbol{\Theta}, \boldsymbol{\Theta}_0]$, $[\boldsymbol{X}\boldsymbol{\Theta}, \boldsymbol{y}]$, respectively. With the definition

$$\mathcal{Z}_n(\mathscr{A}, \mathscr{B}) := \left\{ \text{local minima of } \hat{R}_n(\boldsymbol{\Theta}) \text{ s.t. } \hat{\mu}_{\sqrt{d}[\boldsymbol{\Theta},\boldsymbol{\Theta}_0]} \in \mathscr{A}, \hat{\nu}_{[\boldsymbol{X}\boldsymbol{\Theta},\boldsymbol{y}]} \in \mathscr{B} \right\},$$

for appropriate $\mathscr{A}, \mathscr{B}$, we obtain a bound of the form

$$\lim_{n,d\to\infty} \frac{1}{n} \log \mathbb{E}\left[|\mathcal{Z}_n(\mathscr{A}, \mathscr{B})|\right] \le -\inf_{\mu\in\mathscr{A},\nu\in\mathscr{B}} \Phi(\mu, \nu)$$

for some $\Phi : \mathscr{P}(\mathbb{R}^{k+k_0}) \times \mathscr{P}(\mathbb{R}^{k+q}) \to \mathbb{R}$.

Via Markov's inequality, this bound allows to determine the positions of these minimizers (with exponential deviation bounds) and hence derive sharp asymptotics on the estimation and prediction error: for a given test function $\psi : \mathbb{R}^k \times \mathbb{R}^{k_0} \to \mathbb{R}$ and any local minimizer $\hat{\boldsymbol{\Theta}}$, (with $\hat{\boldsymbol{\Theta}}_i$, $\boldsymbol{\Theta}_{0\,i}$ the i-th rows of $\hat{\boldsymbol{\Theta}}$, $\boldsymbol{\Theta}_0$), we have

$$\mathbb{P}\left\{ \frac{1}{d}\sum_{i=1}^{d} \psi\big(\sqrt{d}\hat{\boldsymbol{\Theta}}_i, \sqrt{d}\boldsymbol{\Theta}_{0\,i}\big) \in I \right\} \le \exp\left\{ -n \inf_{\mu\in\mathscr{A}(I,\psi),\nu} \Phi(\mu, \nu) + o(n) \right\},$$

where $\mathscr{A}(I, \psi) := \{\mu : \int \psi(\boldsymbol{t}, \boldsymbol{t}_0)\mu(\mathrm{d}\boldsymbol{t}, \mathrm{d}\boldsymbol{t}_0) \in I\}$.

We apply our characterization to convex losses, where high-dimensional asymptotics were not (in general) rigorously established for $k \ge 2$. In this setting, we show that Φ satisfies

$$\Phi(\mu, \nu) \ge 0 \quad \forall \mu, \nu \qquad \text{with} \qquad \Phi(\mu, \nu) = 0 \quad \Leftrightarrow \quad (\mu, \nu) = (\mu_\star, \nu_\star),$$

for some $\mu_\star, \nu_\star$. This implies that $\hat{\mu}_{\sqrt{d}[\boldsymbol{\Theta},\boldsymbol{\Theta}_0]} \overset{w}{\Rightarrow} \mu_\star$ and $\hat{\nu}_{[\boldsymbol{X}\boldsymbol{\Theta},\boldsymbol{y}]} \overset{w}{\Rightarrow} \nu_\star$. We derive a set of fixed point equations characterizing $\mu_\star, \nu_\star$, which describe the high-dimensional asymptotics of ERM, and further show that these fixed point equations correspond to the stationarity conditions of an infinite-dimensional, deterministic convex problem. This description unifies many of the ones previously obtained for ERM under different settings, and rigorously establishes it for some settings where no rigorous proof was available.

T06

Dissecting the Interplay of Attention Paths in a Statistical Mechanics Theory of Transformers

Lorenzo Tiberi [1,2] , **Francesca Mignacco** [3,4] , **Kazuki Iries** [1,2] , **and Haim Sompolinsky** [1,2,5]

[1] *Center for Brain Science, Harvard University, Cambridge, MA, USA*
[2] *Kempner Institute for the Study of Natural and Artificial Intelligence, Harvard University, Cambridge, MA, USA*
[3] *Graduate Center, City University of New York, NY, USA*
[4] *Joseph Henry Laboratories of Physics, Princeton University, NJ, USA*
[5] *Edmond and Lily Safra Center for Brain Sciences, Hebrew University of Jerusalem, Jerusalem, Israel*

Despite the remarkable success of Transformers across a range of state-of-the-art machine learning tasks, theoretical characterizations accounting for their impressive performance remain sparse. To enable tractable analysis, theoretical studies often need to rely on significant architectural simplifications, such as limiting models to a single attention head, a single layer, or both. Consequently, an important feature of Transformers remains to a large extent undercharacterized: the interplay between multiple attention heads across multiple layers.
We present recent work [1] addressing this theoretical gap. We consider a deep multi-head self-attention network, that is closely related to Transformers, yet analytically tractable. By making some simplifying assumptions, such as linearity in the value weights and fixed pre-trained query and key weights, our theory is able to characterize a network with an arbitrary number of layers and attention heads.
Using backpropagating kernel renormalization techniques [2], we develop a theory of Bayesian learning in the model, deriving exact equations for the network's predictor statistics under the finite-width thermodynamic limit, i.e. $N, P \to \infty$, $\alpha = P/N = \mathcal{O}(1)$, where N is the network width and P is the number of training examples. Importantly, we go beyond previous results obtained in the Gaussian process limit [3], in which the number $P = \mathcal{O}(1)$ of examples is too small for the network weights to develop data-driven structures.
Our theory relates the network's generalization performance to its learned interplay between "attention paths", defined as information pathways through different attention heads across layers. We show that the network's mean predictor is expressed as a sum of independent kernels, each one corresponding to a different pair of "attention paths". These kernels are weighted according to a "task-relevant path combination" mechanism that aligns the total kernel with the task labels, improving generalization performance.
The kernel weighting is performed by a path-by-path matrix predicted by our theory, which we are able to express as an empirically measurable function of the network weights. This allows for a qualitative transfer of our insights to more complex models; for example, one trained via gradient descent with no fixing of the query and key weights. As an illustration, we demonstrate an efficient size reduction of such a model, by pruning those attention heads that are deemed less relevant by our theory.
Experiments confirm our findings on synthetic and real-world sequence classification tasks.

[1] L. Tiberi, F. Mignacco, K. Irie, H. Sompolinsky, Adv. Neural Inf. Process. Syst. **37**, 72710(2024).
[2] Q. Li, H. Sompolinsky, Phys. Rev. X **11**, 031059 (2021).
[3] J. Hron, Y. Bahri, J. Sohl-Dickstein, R. Novak, Proc. Mach. Learn. Res. **119**, 4376 (2020).

T07

Memorization and Generalization in Generative Diffusion

Raphaël Urfin[1], **Tony Bonnaire**[1], **Giulio Biroli** [1] **and Marc Mézard**[2]

[1]*Laboratoire de Physique de l'Ecole Normale Supérieure, ENS, Université PSL, CNRS, Sorbonne Université, Université Paris Cité, F-75005 Paris, France*
[2]*Department of Computing Sciences, Bocconi University, Milano, Italy*

Diffusion models are a class of generative models that have achieved state-of-the-art performance in a variety of tasks such as audio and image generation. In order to generate new samples from a target distribution, one has to learn the score function [1, 2]. This can be achieved by gradually adding noise to the data points and by leveraging neural network techniques. However an infinitely expressive neural network would learn the empirical score which memorizes the dataset i.e. it only generates data points used in the training [3]. This is not the case in practice where models managed to generalize i.e. they generate new samples from the target distribution, provided the dataset is large enough [4]. To answer this conundrum, we study the properties of the loss landscape and the training dynamics with numerical experiments and simple analytically tractable models.

[1] Song, Yang and Ermon, Stefano, *Generative Modeling by Estimating Gradients of the Data Distribution*, *Advances in Neural Information Processing Systems* (2019).
[2] Yang, Song and Jascha, Sohl-Dickstein and Diederik, P Kingma and Abhishek, Kumar and Stefano, Ermon and Ben, Poole, *Score-Based Generative Modeling through Stochastic Differential Equations*, *International Conference on Learning Representations* (2021).
[3] Biroli, Giulio and Bonnaire, Tony and de Bortoli, Valentin and Mézard, Marc, *Dynamical regimes of diffusion models*, *Nature Communications* (2024)
[4] Zahra Kadkhodaie and Florentin Guth and Eero P Simoncelli and Stéphane Mallat, *Generalization in diffusion models arises from geometry-adaptive harmonic representations*, *The Twelfth International Conference on Learning Representations*, (2024)

T08

The Rich and the Simple: On the Implicit Bias of Adam and SGD

Bhavya Vasudeva[1], **Jung Whan Lee**[1], **Vatsal Sharan**[1], **and Mahdi Soltanolkotabi**[1]

[1]*(Presenting author underlined) University of Southern California*

Adam is the de facto optimization algorithm for several deep learning applications, but an understanding of its implicit bias and how it differs from other algorithms, particularly standard first-order methods such as (stochastic) gradient descent (GD), remains limited. In practice, neural networks trained with SGD are known to exhibit simplicity bias — a tendency to find simple models. In contrast, we show that Adam is more resistant to such simplicity bias. To demystify this phenomenon, in this paper, we investigate the differences in the implicit biases of Adam and GD when training two-layer ReLU neural networks on a binary classification task involving synthetic data with Gaussian clusters. We find that GD exhibits a simplicity bias, resulting in a linear decision boundary with a suboptimal margin, whereas Adam leads to much richer and more diverse features, producing a nonlinear boundary that is closer to the Bayes optimal predictor. This richer decision boundary also allows Adam to achieve higher test accuracy both in-distribution and under certain distribution shifts. We theoretically prove these results by analyzing the population gradients. To corroborate our theoretical findings, we present empirical results showing that this property of Adam leads to superior generalization under distributional shifts across datasets where neural networks trained with SGD are known to show simplicity bias, as well as benchmark datasets for subgroup robustness.

T09

Training Dynamics of In-Context Learning in Linear Attention

Yedi Zhang[1], **Aaditya K. Singh**[1], **Peter E. Latham**[1,*], **and Andrew Saxe**[1,2,*]

[1]*Gatsby Computational Neuroscience Unit, University College London*
[2]*Sainsbury Wellcome Centre, University College London*
[*]*Co-senior authors*

While attention-based models have demonstrated the remarkable ability of in-context learning [2], the theoretical understanding of how these models acquired this ability through gradient descent training is still preliminary. Towards answering this question, we study the gradient descent dynamics of multi-head linear self-attention trained for in-context linear regression. We examine two parametrizations of linear self-attention: one with the key and query weights merged as a single matrix (common in theoretical studies), and one with separate key and query matrices (closer to practical settings). For the merged parametrization, we show the training dynamics has two fixed points and the loss trajectory exhibits a single, abrupt drop. We derive an analytical time-course solution for a certain class of datasets and initialization. For the separate parametrization, we show the training dynamics has exponentially many fixed points and the loss exhibits saddle-to-saddle dynamics, which we reduce to scalar ordinary differential equations. During training, the model implements principal component regression in context with the number of principal components increasing over training time. Overall, we provide a theoretical description of how in-context learning abilities evolve during gradient descent training of linear attention, revealing abrupt acquisition or progressive improvements depending on how the key and query are parametrized.

[1] Zhang, Yedi, et al. "Training Dynamics of In-Context Learning in Linear Attention." arXiv preprint arXiv:2501.16265 (2025).
[2] Brown, Tom, et al. "Language models are few-shot learners." Advances in neural information processing systems 33 (2020): 1877-1901.