SOE 14: Focus Session: Physics of AI I (joint session SOE/DY)

The focus session is organized by Claudius Gros (Goethe-Universität Franfurt), Moritz Helias (Forschungszentrum Jülich), Peter Sollich (Georg-August-Universität Göttingen)

This focus session brings together experts in the field of physics-inspired theory of machine learning and artificial intelligence, who aim to supplement the engineering-driven success of AI by a principled theory of neural information processing. Contributions will address how statistical and dynamical perspectives explain learning in modern AI systems and how these insights support interpretability as well as prediction of performance, generalization, and the required resources.

Time: Thursday 9:30–11:15 Location: GÖR/0226

Invited Talk SOE 14.1 Thu 9:30 GÖR/0226

Generative AI and diffusion models: a statistical physics approach — •Giulio Biroli — Ecole Normale Superieure, Paris, France

Generative AI represents a groundbreaking development within the broader *Machine Learning Revolution,* significantly influencing technology, science, and society. In this colloquium, I will focus on the state-of-the-art *diffusion models*, which are currently used to generate images, videos, and sounds. They are very fascinating algorithms for physicists, as they are very much connected to concepts from stochastic thermodynamics, particularly time-reversed Langevin dynamics. These diffusion models start from a simple white noise input and make it evolve through a Langevin process to generate complex outputs such as images, videos, and sounds. I will show that statistical physics provides principles and methods to characterise this generation process. Specifically, I will discuss how phenomena such as the transition from memorization to generalization and the emergence of features can be understood through the lens of symmetry breaking, phase transitions, slow dynamics, and methods used to study disordered systems.

SOE 14.2 Thu 10:00 GÖR/0226

Statistical Physics of Classifier-free Diffusion Guidance — •Enrico Ventura[1], Beatrice Achilli[1], Carlo Lucibello[1], and Luca Ambrogioni[2] — [1]Bocconi University, Milan, Italy — [2]Radboud University, Nijmegen, The Netherlands

Classifier-free Guidance (CFG) is a simple yet effective technique that helps diffusion models better follow a user's prompt. By combining standard unconditional diffusion with diffusion conditioned on a specific class of the data, it steers generation toward samples (e.g. images, videos or text) that more clearly reflect the intended content. We propose a description of the sampling dynamics of a diffusion model under CFG based on the statistical mechanics of disordered systems. Specifically, we study the time-dependent transformation of the diffusion potential providing a quantitative prediction of the way a complex target distribution is deformed to improve data generation. Moreover, we leverage our results to propose alternative theory-based guidance schedules that enhance such beneficial effects.

SOE 14.3 Thu 10:15 GÖR/0226

Fundamental operating regimes, hyper-parameter fine-tuning and glassiness: towards an interpretable replica-theory for trained restricted Boltzmann machines — •Alberto Fachechi[1], Elena Agliari[1], Miriam Aquaro[1], Anthony Coolen[2], and Menno Mulder[2] — [1]Department of Mathematics, Sapienza University of Roma, P. le A. Moro 5, 00185 Roma, Italy — [2]Theoretical Biophysics, DCN Donders Institute, Faculty of Science, Radboud University, 6525 AJ Nijmegen, The Netherlands

Since the seminal work by Amit, Gutfreund and Sompolinsky, statistical mechanics of spin-glasses with structural disorder has acquired a crucial role in theoretical investigations of artificial neural networks, as it enables the representation of their generalization and information processing capabilities as phases within the space of parameters. We study the relaxation towards equilibrium of the training procedure of restricted Boltzmann machines with a binary visible layer and a Gaussian hidden layer with an unlabelled dataset consisting of noisy realizations of a single ground pattern. We develop a statistical mechanics framework to describe the network generative capabilities by exploiting replica theory. We outline the effective control parameters (e.g., the relative number of weights to be trained, the regularization parameter), whose tuning can yield qualitatively different operative regimes. We also provide analytical and numerical evidence for the existence of a sub-region in the space of the hyperparameters where replica-symmetry breaking occurs.

SOE 14.4 Thu 10:30 GÖR/0226

Mirror, Mirror of the Flow: How Does Regularization Shape Implicit Bias? — •Tom Jacobs, Chao Zhou, and Rebekka Burkholz — CISPA Helmholtz Center, Saarbrucken, Germany

Implicit bias plays an important role in explaining how overparameterized models generalize well. Explicit regularization like weight decay is often employed in addition to prevent overfitting. While both concepts have been studied separately, in practice, they often act in tandem. Understanding their interplay is key to controlling the shape and strength of implicit bias, as it can be modified by explicit regularization. To this end, we incorporate explicit regularization into the mirror flow framework and analyze its lasting effects on the geometry of the training dynamics, covering three distinct effects: positional bias, type of bias, and range shrinking. The mirror flow framework relies on Noether style parameter symmetry preservation, the regularization controls them. Our analytical approach encompasses a broad class of problems, including sparse coding, matrix sensing, single-layer attention, and LoRA, for which we demonstrate the utility of our insights. To exploit the lasting effect of regularization and highlight the potential benefit of dynamic weight decay schedules, we propose to switch off weight decay during training, which can improve generalization, as we demonstrate in experiments.

SOE 14.5 Thu 10:45 GÖR/0226

Generalization performance of narrow one-hidden layer networks in the teacher-student setting — Rodrigo Pérez Ortiz[1], •Gibbs Nwemadji[2], Jean Barbier[3], Federica Gerace[1], Alessandro Ingrosso[4], Clarissa Lauditi[5], and Enrico Malatesta[6] — [1]Alma Mater Studiorum * Università di Bologna (Unibo), Bologna, Italy — [2]International School of Advanced Studies (SISSA), Trieste, Italy — [3]The Abdus Salam International Centre for Theoretical Physics, Trieste, Italy — [4]Radboud University, Nijmegen, The Netherlands — [5]Harvard University, Cambridge, US — [6]Bocconi University, Milano, Italy

Generalization on simple input-output distributions is best studied in the teacher-student setting, but fully connected one-hidden-layer networks with generic activations still lack a complete theory. We develop such a framework for networks with a large but finite number of hidden neurons, using statistical-physics tools to obtain closed-form predictions for both Bayesian and ERM estimators through a few summary statistics. We also identify a specialization transition when the sample size matches the number of parameters. The resulting theory accurately predicts generalization errors for networks trained with Langevin dynamics or standard full-batch gradient descent.

SOE 14.6 Thu 11:00 GÖR/0226

Testing generalization through tiny task switching frameworks — •Daniel Henrik Nevermann and Claudius Gros — Institut für Theoretische Physik, Goethe-Universität Frankfurt, Deutschland

With an ever-growing interest in advancing the performance and efficiency of large language models (LLMs), and therein particularly the transformer architecture, the need for tiny testing frameworks is pressing, as many researchers cannot afford to train models on large GPU clusters. We here propose a tiny testing framework, extending the recently published IARC task switching framework, that despite being trivial to implement offers suitable complexity to be non-trivial to learn for small scale transformer models with a few million parameters or less. Beyond model benchmarking, the framework is also suitable for probing phenomena relevant to problems arising in physics of AI, where controlled, interpretable testbeds are essential. The proposed training and evaluation scheme relies on integer sequences to be predicted by the model. These integer sequences are generated by simple deterministic tasks designed to abstract typical challenges arising in natural language processing, such as short and long range correlations, or context awareness. Within the sequences, tasks are randomly switched, where a switch is indicated by a control token. An important quality of LLMs is the ability to generalize at inference time. We here extend the existing task switching framework with new tasks able to probe models generalization capacities in a tiny, yet meaningful manner.