1st Symposium on

Machine Learning for Theories and Theories of Machine Learning

29/9/24 — 3/10/24

Rovinj, Croatia

University of Basel

European Research Council
Established by the European Commission

University of Basel

1st Symposium on **Machine Learning for Theories and Theories of Machine Learning**

29/9/24 — 3/10/24

Rovinj, Croatia

Agenda

Sunday, 29 September

6:30 pm	Welcome drink and Reception	

Monday, 30 September

8:20 am	Opening remarks	Ivan Dokmanić
8:30 am	Synthetic Data – Friend or Foe in the Age of Scaling?	Julia Kempe
9:20 am	Asymptotic theory of in-context learning by linear attention	Yue M. Lu
10:10 am	*Coffee/juice break*	
10:40 am	How Feature Learning Can Improve Neural Scaling Laws?	Cengiz Pehlevan
11:30 am	TBD	Maria Brbić
12:20 pm	*Lunch*	
2:30 pm	Mapping properties of neural networks and inverse problems	Matti Lassas
3:20 pm	*Coffee/juice break*	

3:30 pm	TBD	Rava da Silveira
4:20 pm	A formal perspective on language modeling	Ryan Cotterell
5:10 pm	*End of the day*	

Tuesday, 1 October

8:30 am	Generative modeling with flows and diffusions	Eric Vanden-Eijnden
9:20 am	Approximation and Generalisation by Score Diffusion with the Renormalisation Group	Stéphane Mallat
10:10 am	*Coffee/juice break*	
10:40 am	Speeding up gradient flows on probability measure space	Qin Li
11:30 am	Towards a sharp non-asymptotic theory for structured random matrices (and tensors)	Afonso Bandeira
12:20 pm	*Lunch*	
2:30 pm	Learning higher-order data correlations with neural networks, efficiently	Sebastian Goldt
3:20 pm	*Coffee/juice break*	
3:30 pm	Optimization, Robustness and Attention in Deep Learning: Insights from Random and NTK Feature Models	Marco Mondelli
4:20 pm	Information geometry and the hierarchy of effective theories in physics, biology, and beyond	Mark Transtrum
5:10 pm	*End of the day*	

Wednesday, 2 October

8:30 am	Transformers are universal in-context learners	Maarten de Hoop

9:20 am	Machine Learning for Scientific Discovery, with Examples in Fluid Mechanics	Steven Brunton
10:10 am	*Coffee/juice break*	
10:40 am	Model Selection And Ensembling When There Are More Parameters Than Data	Michael Mahoney
11:30 am	Avoiding representational collapse in non-contrastive self-supervised learning - Lessons from neurobiology	Friedemann Zenke
12:20 pm	*Lunch*	
2:30 pm	TBD	Nicolas Flammarion
3:20 pm	*Coffee/juice break*	
3:30 pm	Learning earthquake displacement	Paul Johnson
4:20 pm	What does the neuron do? A new model for neuroscience and AI	Mitya Chklovskii
5:10 pm	*End of the day*	

Thursday, 3 October

8:30 am	Emergent mechanisms in transformers: A sample complexity and an architectural perspective	Freya Behrens
9:20 am	Bridging Physics and Computation through the Neuromorphic Intermediate Representation	Steven Abreu
10:10 am	A spring--block theory of feature learning in deep neural networks	Cheng Shi

Abstracts — Monday, 30 September

8:30 am **Julia Kempe** — New York University

Synthetic Data – Friend or Foe in the Age of Scaling

As AI model size grows, neural **scaling laws** have become a crucial tool to predict the improvements of large models when increasing capacity and the size of original (human or natural) training data. Yet, the widespread use of popular models means that the ecosystem of online data and text will co-evolve to progressively contain increased amounts of synthesized data.In this talk we ask: **How will the scaling laws change in the inevitable regime where synthetic data makes its way into the training corpus?** Will future models, still improve, or be doomed to degenerate up to total **(model) collapse**? We develop a theoretical framework of model collapse through the lens of scaling laws. We discover a wide range of decay phenomena, analyzing loss of scaling, shifted scaling with number of generations, the "un-learning" of skills, and grokking when mixing human and synthesized data. Our theory is validated by large-scale experiments with a transformer on an arithmetic task and text generation using the LLM Llama2. We also propose solutions to circumvent degradation in learning by pruning the generated data.

9:20 am **Yue M. Lu** — Harvard University

Asymptotic theory of in-context learning by linear attention

Transformers have a remarkable ability to learn and execute tasks based on examples provided within the input itself, without explicit prior training. It has been argued that this capability, known as in-context learning (ICL), is a cornerstone of Transformers' success, yet questions about the necessary sample complexity, pretraining task diversity, and context length for successful ICL remain unresolved. Here, we provide a precise answer to these questions in an exactly solvable model of ICL of a linear regression task by linear attention. We derive sharp asymptotics for the learning curve in a phenomenologically-rich scaling regime where the token dimension is taken to infinity; the context length and pretraining task diversity scale proportionally with the token dimension; and the number of pretraining examples scales quadratically. We demonstrate a double-descent learning curve with increasing pretraining examples, and uncover a phase transition in the model's behavior between low and high task diversity regimes: In the low diversity regime, the model tends toward memorization of training tasks, whereas in the high diversity regime, it achieves genuine in-context learning and generalization beyond the scope of pretrained tasks. These theoretical insights are empirically validated through experiments with both linear attention and full nonlinear Transformer architectures.

Joint work with Mary Letey, Jacob Zavatone-Veth, Anindita Maiti, and Cengiz Pehlevan. https://arxiv.org/abs/2405.11751.

The main technical ingredient of our analysis is characterizing the spectral properties of a sample covariance matrix formed by tensorized versions of random vectors. Related

but simpler models of random matrices have been studied in recent work on kernel random matrices in polynomial scaling regimes. In the talk, I will discuss some of the key ideas in analyzing these matrices.

Joint work with Sofiia Dubova, Benjamin McKenna, and Horng-Tzer Yau https://arxiv.org/abs/2310.18280.

10:40 am **Cengiz Pehlevan** — Harvard University

How Feature Learning Can Improve Neural Scaling Laws

We develop a simple solveable model of neural scaling laws beyond the lazy learning regime. Theoretical analysis of this model provides scaling predictions for the model size, training time and total amount of available data.

11:30 am **Maria Brbić** — EPFL

2:30 pm **Matti Lassas** — University of Helsinki

Mapping properties of neural networks and inverse problems

We will consider mapping properties of neural networks and neural operators which are infinite dimensional generalizations of neural networks. In particular, we consider the injectivity of neural networks and universal approximation property of injective neural networks. In addition, we study approximation of probability measures using neural networks that are compositions of invertible flow networks and injective layers and present applications in inverse problems. We also discuss how continuous deformations in function spaces (e.g. functions $u:[0,1]^2\to \mathbb{R}$ that model images) can be approximated by finite dimensional models.The talk is based on collaboration with Maarten de Hoop, Ivan Dokmanic, Takashi Furuya, Konik Kothari, Pekka Pankka and Michael Puthawala.

3:30 pm **Rava da Silveira** — CNRS / ENS / IOB

4:20 pm **Ryan Cotterell** — ETH Zurich

A formal perspective on language modeling

Language models—especially the large ones—are all the rage. And, for what will surely be one of only a few times in history, my field, natural language processing, is the center of world attention. Indeed, there is nearly a daily stream of articles in the popular press on the most recent advances in language modeling technology. In contrast to most of these articles (and most other talks on the topic), this tutorial-style presentation is not about forward progress in the area. Instead, I am going to take a step back and ask simple questions about the nature of language modeling itself. We will start with the

most basic of questions: From a mathematical perspective, what is a language model? Next, the talk will turn philosophical. With all the talk of artificial general intelligence, what can theory of computation bring to bear on the computational power of language models? The talk will conclude with a statement of several recent theorems proven by my research group, the highlight of which is that no Transformer-based language model is Turing complete and, thus, we should be careful about labeling such language models, e.g., GPT-4, as general-purpose reasoners.

Abstracts — Tuesday, 1 October

8:30 am **Eric Vanden-Eijnden** — New York University

Generative modeling with flows and diffusions

Generative models based on dynamical transport have recently led to significant advances in unsupervised learning. At mathematical level, these models are primarily designed around the construction of a map between two probability distributions that transform samples from the first into samples from the second. While these methods were first introduced in the context of image generation, they have found a wide range of applications, including in scientific computing where they offer interesting ways to reconsider complex problems once thought intractable because of the curse of dimensionality. In this talk, I will discuss the mathematical underpinning of generative models based on flows and diffusions, and show how a better understanding of their inner workings can help improve their design. These results indicate how to structure the transport to best reach complex target distributions while maintaining computational efficiency, both at learning and sampling stages. I will also discuss applications of generative AI in scientific computing, in particular in the context of Monte Carlo sampling, with applications to statistical mechanics and Bayesian inference, as well as probabilistic forecasting, with application to fluid dynamics and atmosphere/ocean science

9:20 am **Stéphane Mallat** — Collège de France/ENS

Approximation and Generalisation by Score Diffusion with the Renormalisation Group

Score based diffusions generate impressive models of images, sounds and complex physical systems. Are they generalising or memorising ? How can deep network estimate high-dimensional scores without curse of dimensionality ? This talk adresses both questions and concentrates on structured estimations of scores. It is related to the renormalisation group in statistical physics and to harmonic analysis models based on sparsity.Defining functional classes of complex fields such as turbulences is an outstanding problem that will be discussed.

10:40 am **Qin Li** — University of Wisconsin

Speeding up gradient flows on probability measure space

In the past decade, there has been a significant shift in the types of mathematical objects under investigation, moving from vectors and matrices in the Euclidean spaces to functions residing in Hilbert spaces, and ultimately extending to probability measures within the probability measure space. Many questions that were originally posed in the context of linear function spaces are now being revisited in the realm of probability measures. One such question is to the efficiently find a probability measure that minimizes a given objective functional.In Euclidean space, we devised optimization techniques such as gradient descent and introduced momentum-based methods to accelerate its convergence. Now, the question arises: Can we employ analogous strategies to expedite convergence within the probability measure space?
In this presentation, we provide an affirmative answer to this question. Specifically, we present a series of momentum-inspired acceleration method under the framework of Hamiltonian flow, and we prove the new class of method can achieve arbitrary high-order of convergence. This opens the door of developing methods beyond standard gradient flow.

11:30 am **Afonso Bandeira** — ETH Zurich

Towards a sharp non-asymptotic theory for structured random matrices (and tensors)

Matrix Concentration inequalities such as Matrix Bernstein inequality (Oliveira and Tropp) have played an important role in many areas of pure and applied mathematics. These inequalities are intimately related to the celebrated noncommutative Khintchine inequality of Lust-Piquard and Pisier. While these tend to be optimal when the underlying matrices are commutative, they are known to be sub-optimal in several other instances. Recently, we have leveraged ideas from Free Probability to fully remove the sub-optimal dimensional dependencies in these inequalities in a range of instances, yielding sharp bounds in many settings of interest. In this talk I will describe these results, some of the recent and ongoing work that it has sparked, and several open problems.
Includes joint work with: March Boedihardjo (MSU); Ramon van Handel and Giorgio Cipolloni (Princeton); Petar Nizic-Nikolac, Anastasia Kireeva, Kevin Lucca, and Dominik Schroder (ETH); Xinmeng Zeng (Stanford); Dustin Mixon (OSU); Dmitriy Kunisky (Johns Hopkins); Thomas Rothvoss (U Washington); Haotian Jiang (U Chicago); Sivakanth Gopi (MSR).

2:30 pm **Sebastian Goldt** — SISSA

Learning higher-order data correlations with neural networks, efficiently

Neural networks excel at finding patterns in their data -- but which patterns do they extract, and how do they extract them efficiently? I will discuss two recent lines of work that look at this problem in the context of computer vision and language modelling. First, I will describe a mechanism by which neural networks can learn from the higher- order correlations of images -- a computationally hard task! -- efficiently by exploiting correlations in the latent space of the inputs. In the second part, I will discuss how transformers learn increasingly complex distributions of their inputs.

3:30 pm **Marco Mondelli** — IST Austria

Optimization, Robustness and Attention in Deep Learning: Insights from Random and NTK Feature Models

A recent line of work has analyzed the properties of deep learning models through the lens of Random Features (RF) and the Neural Tangent Kernel (NTK). In this talk, I will show how concentration bounds on RF and NTK maps provide insights on (i) the optimization of the network via gradient descent, (ii) its adversarial robustness, and (iii) the success of attention-based architectures, such as transformers.I will start by proving tight bounds on the smallest eigenvalue of the NTK for deep neural networks with minimum over-parameterization. This implies that the network optimized by gradient descent interpolates the training dataset (i.e., reaches 0 training loss), as soon as the number of parameters is information-theoretically optimal. Next, I will focus on the robustness of the interpolating solution. A thought-provoking paper by Bubeck and Sellke has proposed a “universal law of robustness”: interpolating smoothly the data necessarily requires many more parameters than simple memorization. By providing sharp bounds on RF and NTK models, I will show that, while some random feature models cannot be robust (regardless of the over-parameterization), NTK features are able to saturate the universal law of robustness, thus addressing a conjecture by Bubeck, Li and Nagaraj. Finally, I will consider attention-based architectures, showing that random attention features are sensitive to a change of a single word in the context, as expected from a model suitable for NLP tasks. In contrast, the sensitivity of random features decays with the length of the context. This property translates into generalization bounds: due to their low word sensitivity, random features provably cannot learn to distinguish between two sentences that differ only in a single word. In contrast, due to their high word sensitivity, random attention features have higher generalization capabilities.

4:20 pm **Mark Transtrum** — Brigham Young University

Information geometry and the hierarchy of effective theories in physics, biology, and beyond

The success of science is due in large part to the hierarchical nature of physical theories. These effective theories model natural phenomena as if the physics at macroscopic length scales were almost independent of the underlying, shorter-length-scale details. I interpret this hierarchy in terms of parameter identifiability. Parameters associated with microscopic degrees of freedom are usually unidentifiable as quantified by the Fisher Information Matrix. I then apply an information geometric approach in which a microscopic, mechanistic model is interpreted as a manifold of predictions in data space. Model manifolds are often characterized by a hierarchy of boundaries--faces, edges, corners, hyper-corners, etc. These boundaries correspond to reduced-order models, leading to a model reduction technique known as the Manifold Boundary Approximation Method. In this way, effective models can be systematically derived from microscopic first principles for a variety of complex systems in physics, biology, and other fields. I conclude by discussing applications of order theory to reason about the hierarchical structures in models and theories.

University of Basel

Abstracts — Wednesday, 2 October

8:30 am **Maarten de Hoop** — Rice University

Transformers are universal in-context learners

Transformers are deep architectures that define "in-context mappings" which enable predicting new so-called tokens based on a given set of tokens as the input data. We present the ability of these architectures to handle an arbitrarily large number of context tokens. To mathematically and uniformly address their expressivity, we consider the case that the mappings are conditioned on a context represented by a probability distribution of tokens. The relevant notion of smoothness then corresponds to continuity in terms of the Wasserstein distance between such contexts. We demonstrate that deep transformers are universal and can approximate continuous in-context mappings to arbitrary precision, uniformly over compact token domains. A key aspect of our results is that for a fixed precision, a single transformer can operate on an arbitrary (even infinite) number of tokens. Additionally, it operates with a fixed embedding dimension of tokens (this dimension does not increase with precision) and a fixed number of heads (proportional to the dimension). The use of multilayer perceptrons between multi-head attention layers is also explicitly controlled. We consider both the un-masked and masked settings; in the masked setting, attentions are no longer permutation equivariant as the masking imposes a causality constraint.Joint work with Takashi Furuya and Gabriel Peyré.

9:20 am **Steven Brunton** — University of Washington

Machine Learning for Scientific Discovery, with Examples in Fluid Mechanics

Accurate and efficient nonlinear dynamical systems models are essential understand, predict, estimate, and control complex natural and engineered systems. In this talk, I will explore how machine learning may be used to develop these models purely from measurement data. We explore the sparse identification of nonlinear dynamics (SINDy) algorithm, which identifies a minimal dynamical system model that balances model complexity with accuracy, avoiding overfitting. This approach tends to promote models that are interpretable and generalizable, capturing the essential “physics” of the system. We also discuss the importance of learning effective coordinate systems in which the dynamics may be expected to be sparse. This sparse modeling approach will be demonstrated on a range of challenging modeling problems, for example in fluid dynamics. Because fluid dynamics is central to transportation, health, and defense systems, we will emphasize the importance of machine learning solutions that are interpretable, explainable, generalizable, and that respect known physics.

10:40 am **Michael Mahoney** — UC Berkeley

Model Selection And Ensembling When There Are More Parameters Than Data

Despite years of empirical success with deep learning for many large-scale problems, existing theoretical frameworks fail to explain many of the most successful heuristics used by practitioners. The primary weakness most approaches encounter is a reliance on the typical large data regime, which neural networks often do not operate in due to their large size. To overcome this issue, I will describe how for any overparameterized (high-dimensional) model, there exists a dual underparameterized (low-dimensional) model that possesses the same marginal likelihood, establishing a form of Bayesian duality. Applying classical methods to this dual model reveals the Interpolating Information Criterion, a measure of model quality that is consistent with current deep learning heuristics. I will also describe how, in many modern machine learning settings, the benefits of ensembling are less ubiquitous and less obvious than classically. Theoretically, we prove simple new results relating the ensemble improvement rate (a measure of how much ensembling decreases the error rate versus a single model, on a relative scale) to the disagreement-error ratio. Empirically, the predictions made by our theory hold, and we identify practical scenarios where ensembling does and does not result in large performance improvements. Perhaps most notably, we demonstrate a distinct difference in behavior between interpolating models (popular in current practice) and non-interpolating models (such as tree-based methods, where ensembling is popular), demonstrating that ensembling helps considerably more in the latter case than in the former.

11:30 am **Friedemann Zenke** — FMI Basel

Avoiding representational collapse in non-contrastive self-supervised learning - Lessons from neurobiology

In this talk, I will explain how non-contrastive self-supervised learning (SSL) based on variance regularization relates to classic synaptic plasticity models in neurobiology. To that end, I will introduce Latent Predictive Learning (LPL), a model integrating Hebbian and predictive plasticity to develop unsupervised, disentangled object representations in deep neural networks. I will illustrate how the resulting learning rules explain neuronal selectivity changes in the primate inferotemporal cortex. Moreover, I will show how implicit variance regularization through the learning dynamics of the predictor network avoids collapse in established non-contrastive joint-embedding SSL approaches like BYOL and SimSiam.

2:30 pm **Nicolas Flammarion** — EPFL

3:30 pm **Paul Johnson** — Los Alamos National Laboratory

Learning earthquake displacement

Machine learning algorithms have shown great success in advancing science and technology in seismology for a number of applications. These include applications to earthquake detection, classification of different seismic signals, and predicting displacements and time-to-failure in laboratory shear experiments, slow slip in subduction zones and volcanic earthquakes where we have sufficient training and testing data. In recent work, we apply machine learning as an exploratory data analysis tool to investigate whether, (a) features in the recorded, continuous seismic signals contain information about the fault displacement, as they do in the laboratory; and (b) if precursory information regarding an upcoming failure is contained in the continuous seismic signal. We search for patterns that may not be captured by traditional signal processing techniques. Continuous records of digital seismograms are acquired for analysis from a volcano caldera-collapse earthquake-sequence of >60 magnitude ~5 earthquakes at the Kilauea volcano in Hawaii. Labels are ground displacements recorded by global navigation satellite system (GNSS) stations, used as a proxy for fault displacement. We experiment using supervised learning with shallow architecture gradient boosted trees as well as deep learning methods using transformers. We find the models do an excellent job in predicting contemporaneous fault displacement and there are suggestions that near-future information is contained in the continuous seismic signals.

4:20 pm **Mitya Chklovskii** — Flatiron Institute/NYU Medical Center

What does the neuron do? A new model for neuroscience and AI

Modern Artificial Intelligence (AI) systems, such as ChatGPT, rely on artificial neural networks (ANNs), which are historically inspired by the human brain. Despite this inspiration, the similarity between ANNs and biological neural networks is largely superficial. For instance, the foundational McCulloch-Pitts-Rosenblatt unit of ANNs drastically oversimplifies the complexity of real neurons.Recognizing the intricate temporal dynamics in biological neurons and the ubiquity of feedback loops in natural networks, we suggest reimagining neurons as feedback controllers. A practical implementation of such controllers within biological systems is made feasible by the recently developed Direct Data-Driven Control (DD-DC). We find that DD-DC neuron models can explain various neurophysiological observations, affirming our theory.

Abstracts — Thursday, 3 October

8:30 am **Freya Behrens** — EPFL

Emergent mechanisms in transformers: A sample complexity and an architectural perspective

This talk explores how transformer models are able to implement different problem-solving strategies with the same architecture, and how different architectural choices bias the problem solving strategy. We first show how a dot-product attention layer exhibits a shift from positional to semantic attention as data complexity increases. The

emergence of this phase transition is backed by theoretical analysis using statistical physics backed up by experiments. Next, we examine how small transformers handle counting tasks, revealing two distinct methods: relation-based counting, which relies on efficient attention mixing, and inventory-based counting, which uses more memory and computation. Using ideas from mechanistic interpretability we identify how subtle architectural differences determine which method the model learns. (based on arxiv:2402.03902 and arxiv:2407.11542)

9:20 am **Steven Abreu** — University of Groningen

Bridging Physics and Computation through the Neuromorphic Intermediate Representation

The Neuromorphic Intermediate Representation (NIR) offers a framework to map abstract computational graphs from machine learning into physical substrates, by modeling the underlying computations as continuous-time dynamics on graphs. This approach provides a versatile modeling framework for physics-based computing and emerging hardware platforms such as neuromorphic processors. In this talk, I will present how NIR aligns computation with the underlying physics of neuromorphic devices, enhancing interoperability across different systems. I will discuss our successes in reproducing NIR graphs across various hardware platforms, and explore the theoretical foundations that address challenges like device mismatch and limited observability in continuous-time computation. The talk will conclude with future directions for NIR, particularly in modeling computation in fully analog systems.

10:10 am **Cheng Shi** — University of Basel

A spring--block theory of feature learning in deep neural networks

A central question in deep learning is how deep neural networks (DNNs) learn features. DNN layers progressively collapse data into a regular low-dimensional geometry. This collective effect of nonlinearity, noise, learning rate, width, depth, and numerous other parameters, has eluded first-principles theories which are built from microscopic neuronal dynamics. Here we present a noise--nonlinearity phase diagram that highlights where shallow or deep layers learn features more effectively. We then propose a macroscopic mechanical theory of feature learning that accurately reproduces this phase diagram, offering a clear intuition for why and how some DNNs are ``lazy'' and some are ``active'', and relating the distribution of feature learning over layers with test accuracy.