Mathematisches Forschungsinstitut Oberwolfach

Report No. 4/2025

DOI: 10.4171/OWR/2025/4

Overparametrization, Regularization, Identifiability and Uncertainty in Machine Learning

Organized by
Nicolò Cesa-Bianchi, Milano
Philipp Hennig, Tübingen
Andreas Krause, Zürich
Ulrike von Luxburg, Tübingen

26 January – 31 January 2025

Abstract. In machine learning, a field addressing the extraction of information and structure from finite data with the means of computer science and mathematics, maps from finite-dimensional spaces of data or computations into spaces of higher, or infinite dimensionality are a central theme. The workshop brought together researchers with diverse viewpoints to discuss how different theoretical sub-communities within the field treat the resulting ill-posed operations, and what kind of features of algorithms and models can emerge as a result.

Mathematics Subject Classification (2020): 62-XX, 68-XX.

Introduction by the Organizers

1. Topical Overview

While machine learning (ML) currently enjoys outsized interest by the public as well as by multiple scientific domains, it is still a relatively young field, drawing methods, concepts, and formalisms from older domains. Since its inception, the ML community has struck a productive balance between empirical and applied work on the one hand, and a desire for rigorous theoretical analysis on the other. This has allowed field to identify new, and often disruptive ideas quickly, but to then also develop them into general, efficient, well-understood frameworks. In the early years of this decade, the empirical side of the field has once again taken the lead, making ground-breaking advances in particular in the area of generative modelling from unsupervised data using deep neural architectures (parametrized differentiable functions) to address extremely high-dimensional structured probability distributions. These models, which were elevated with two Nobel prizes in 2024, pose many theoretical questions about their behaviour, limitations, and efficient algorithmic means to train and control them.

Oberwolfach Workshop 4/2025 sought to bring together members of multiple theoretical sub-communities of machine learning, mostly from Europe, to discuss recent advances in their work. We made a deliberate effort to connect people with different perspectives to draw connections across the field, even if this meant that the title of the workshop became a bit unwieldy. For, as the field continues to grow, similar observations are made repeatedly, from different viewpoints and using different technical vocabularies. The central operation of machine learning is that of inference: Mapping from a finite-dimensional space of data, prompts, inputs into a typically much higher – or infinite-dimensional space of answers, hypotheses, or otherwise latent variables. This is principally an ill-posed process, and different sub-communities address the resulting challenges and features differently. A few examples of the resulting questions are as follows:

- From the perspective of **probability theory**, identifiability is associated with uncertainty quantification. But most contemporary deep architectures do not quantify epistemic uncertainty. Several of the talks at the workshop discussed approximate techniques to endow advanced, contemporary deep models with such posterior probability measures, including for models with nonparametric, function-valued outputs. **Numerical methods** commonly used in machine learning (optimization, least-squares, simulation methods, etc.) also have internal discretization leading to misspecification and uncertainty. This error interacts non-trivially with the empirical estimation error caused by finite data. By modelling computational error with probability measures **Probabilistic numerics** envisions a unified notion of uncertainty / error estimation across the methodological stack, from algorithm to model, incorporating data and computation. Several talks covered such methods, and their integrative use within the ML stack.
- Partial monitoring is the most fundamental paradigm for the analysis of **sequential decision-making problems**. In this framework, the performance of an algorithm is typically measured via the regret against the optimal decision in hindsight. The structure of the decision space, the shape of the utility function, the nature of the environment, and the amount of feedback the learner receives after each decision round are the parameters that control the optimal (minimax) regret. While the landscape is relatively well understood for finite decision spaces, the crucial properties of the parameters the determine the minimax rates in **continuous decision spaces** are yet to be clarified. The workshop included talks discussing the

implications of regret minimization in multi-agent settings, such as online market problems.

- The traditional model of learning theory assumes a **stationary** joint probability distribution over a set of observables (random variables), from which both training and test observations are sampled. In practice, however, this model is inappropriate in at least three significant aspects. First, real-world data generating distributions may drift between training and testing, calling for suitable representations of, and metrics for, probability distributions, methods to identify **invariant components**, and learning algorithms incorporating those (out-of-distribution generalization). Some speakers discussed how recently emerging model classes, e.g. for sequential data-types, provide new ways to do so. The set of variables that are jointly observed may be also incomplete and differ from training to testing (out-of-variables generalization). And real-world scenarios will often include agents that **intervene** in a system. We heard talks explaining that, when, and why, it can still be possible to identify hypotheses correctly in such settings under mild assumptions.
- One of the goals of this workshop was investigating questions at the intersection of **reinforcement learning** and **causal discovery**. System identification and model estimation in reinforcement learning (RL) are effectively causal discovery tasks. Yet, so far, the interaction between the causal discovery and theoretical RL communities has been limited. Several speakers discussed how insights from causal discovery can be used to analyse and accelerate system identification and model estimation in reinforcement learning, and to what extent the opportunity to *explore* in RL remedy *identifiability* issues in causal discovery. Another, unusual perspective on the interactive setting we discussed is the social question whether humans affected by algorithmic decisions can deliberately use their role as the source of training data to affect and change the outcome of an algorithmic decision in a manner desirable to them.
- Only few years ago, the success of **deep learning** questioned traditional machine learning theory and the established mechanisms that were thought to enable successful learning and generalization: small function classes, regularization, and convex problems. In statistics, researchers could not understand why the bias-variance trade-off did not seem to matter any more, and in optimization, people were puzzled why simple stochastic gradient descent was so successful on a highly non-convex optimization surface. Since then, a new era of learning theory has unfolded. First important steps and mechanisms have been achieved by now (*benign over-fitting, double descent, robust interpolation*, etc.), but formal results often concern simple model classes (linear problems in high-dimensional spaces). We urgently need to understand how these methods scale. Limit results like the *neural tangent kernel* apply to simple network architectures (few layers, infinitely wide), but we still do not understand the complex architectures

used in practice (why do we need depth in networks? Do transformers actually add anything fundamental beyond scale?). **Modern learning theory** is still in its infancy, so teasing out the right questions to ask and concepts to establish was a high-level goal of our discussions. Several talks at the workshop provided new insights on this end, analysing both the model classes used in modern architectures and their capacity, and the optimization methods used to train these models.

2. Summary of Plenary Discussions

There were two plenary discussions on a set of topics suggested by all participants. The first discussion was introduced by Yishay Mansour, who presented a historical perspective on the development of Artificial Intelligence and Machine Learning, highlighting how the dominating paradigm shifted from symbolic manipulation to data-driven methods. Some of the most notable successes of the field that were mentioned included: statistical learning with sample complexity and computational complexity analyses of learning problems, the online learning model, the connection with optimization (stochastic gradient descent and Tikhonov regularization), probabilistic modeling, causality, and reinforcement learning. Some approaches that did not live up to their expectations were also mentioned. For example: logics, inductive inference, and learning of formal languages. The ensuing discussion concerned the most problematic issues in the current research. For example, how to reduce the growing gap between theory and practice in Machine Learning and how theory can inform the engineering of large-scale systems (like Large Language Models). The final part of the discussion involved emerging issues in the teaching of Machine Learning, especially the tension between the teaching of foundational aspects and the pressure from the job market. The second plenary discussion, introduced by Alberto Bietti, was more technical and mainly focused on the connections between overparameterization, scaling laws, and generalization in deep learning and large language models.

Workshop: Overparametrization, Regularization, Identifiability and Uncertainty in Machine Learning

Table of Contents

Abstracts

Selective induction Heads: How Transformers Select Causal Structures in Context

Nicolas Flammarion
(joint work with Francesco D'Angelo, Francesco Croce)

Transformers have exhibited exceptional capabilities in sequence modelling tasks, leveraging self-attention and in-context learning. Critical to this success are induction heads, attention circuits that enable copying tokens based on their previous occurrences. In this talk, we introduce a novel synthetic framework designed to enable the theoretical analysis of transformers' ability to dynamically handle causal structures. Existing works rely on Markov Chains to study the formation of induction heads, revealing how transformers capture causal dependencies and learn transition probabilities in-context. However, they rely on a fixed causal structure that fails to capture the complexity of natural languages, where the relationship between tokens dynamically changes with context. To this end, our framework varies the causal structure through interleaved Markov chains with different lags while keeping the transition probabilities fixed. This setting unveils the formation of Selective Induction Heads, a new circuit that endows transformers with the ability to select the correct causal structure in-context. We empirically demonstrate that attention-only transformers learn this mechanism to predict the next token by identifying the correct lag and copying the corresponding token from the past. We provide a detailed construction of a 3-layer transformer to implement the selective induction head, and a theoretical analysis proving that this mechanism asymptotically converges to the maximum likelihood solution. Our findings advance the theoretical understanding of how transformers select causal structures, providing new insights into their functioning and interpretability.

References

[1] F. D'Angelo, F. Croce, and N. Flammarion *Selective induction Heads: How Transformers Select Causal Structures in Context*, The Thirteenth International Conference on Learning Representations, 2025.

Recent Trends in Learning Operators

Nicole Mücke
(joint work with Mike Nguyen)

Operator learning extends machine learning principles, originally developed for function estimation from finite-dimensional data, to the estimation of operators from infinite-dimensional data. This paradigm enables the supervised learning of operators between function spaces, offering a natural framework for accelerating scientific computation and discovery. Such a framework can facilitate the development of fast surrogate models that approximate costly existing simulations or enable the discovery of new models consistent with observed data in the absence of first-principles-based models.

We approach this topic from a learning-theoretical perspective, introducing the neural tangent kernel (NTK) regime for two-layer neural operators and analyzing their generalization properties. For early-stopped gradient descent, we derive fast convergence rates that are minimax optimal within the framework of nonparametric regression in reproducing kernel Hilbert spaces (RKHS). Additionally, we provide bounds on the number of hidden neurons and second-stage samples required for generalization.

To justify the NTK regime, we show that any operator approximable by a neural operator can also be approximated by an operator from the RKHS associated with the NTK.

A key application of neural operators is learning surrogate maps for solution operators of partial differential equations (PDEs). To illustrate our theoretical findings, we demonstrate simulations using the standard Poisson equation.

References

[1] Mike Nguyen and Nicole Mücke, *Optimal Convergence Rates for Neural Operators*, arXiv:2412.17518

[2] M. Mollenhauer, N. Mücke, and JT Sullivan, *Learning linear operators: Infinite dimensional regression as a well-behaved non-compact inverse problem*, arXiv:2211.08875

Understanding Aggregate Shap Values

Rhobi Bhattacharjee

The shap explainability methods one of the most widely used local explanation methods. One popular use case is for feature selection where practitioners compute the average (absolute) to eliminate unimportant features despite its widespread use, the theoretical underpinnings of this practice remain unexplored.

In this work we investigate how theoretically sound the idea is. We give sufficient conditions under which a uniformly 0 Shapley value provably implies that the corresponding feature can be safely removed without a loss in performance. Our analyses involve a natural Lie algebra constructed with Shapley value functions that may be of independent interest.

Theoretical Foundations and Optimization of Deep State-Space Models

Antonio Orvieto

Structured state-space models (SSMs) are gaining popularity as effective foundational architectures for sequential data, demonstrating outstanding performance across a diverse set of domains alongside desirable scalability properties. Recent developments show that if the linear recurrence powering SSMs allows for a selectivity mechanism leveraging multiplicative interactions between inputs and hidden states (e.g. Mamba (2), GLA, Griffin..), then the resulting architecture can surpass attention-powered foundation models trained on text in both accuracy and efficiency, at scales of billion parameters. In this talk, we give an introduction to SSM and study their expressivity using tools from Rough Path Theory. We provide a framework for the theoretical analysis of generalized selective SSMs, fully characterizing their expressive power and identifying the gating mechanism as the crucial architectural choice. Our study provides a closed-form description of the expressive powers of modern SSMs, such as Mamba, quantifying theoretically the drastic improvement in performance from the previous generation of models, such as S4. Our theory not only motivates the success of modern selective state-space models but also provides a solid framework to understand the expressive power of future SSM variants. In particular, it suggests cross-channel interactions could play a vital role in future improvements. We will discuss on topics such as optimization of recurrent layers, and outline a few directions for future improvements.

Computation-aware Gaussian Processes: Blurring the Line Between Computation and Inference

Jonathan Wenger

(joint work with Kaiwen Wu, Philipp Hennig, Jacob R. Gardner, Geoff Pleiss, John P. Cunningham)

Inference and model selection in Gaussian processes scales prohibitively with the size of the training dataset, both in time and memory. While many approximations exist, all incur inevitable approximation error. We will introduce a new class of GP approximations for which, surprisingly, this inevitable error can be tractably quantified in the form of computational uncertainty. This enables an explicit tradeoff between computational efficiency and precision. This class of computation-aware GPs extends a range of existing approximations and enjoys strong theoretical guarantees. Experiments demonstrate that our approach can outperform state-of-the-art methods like SGPR, CGGP and SVGP on benchmark regression datasets with up to 1.8M datapoints, while requiring only a single GPU. In summary, we demonstrate how to train Gaussian processes on large-scale datasets without significantly compromising their ability to quantify uncertainty – a fundamental prerequisite for optimal decision-making.

References

[1] J. Wenger, K. Wu, P. Hennig,J. R. Gardner, G. Pleiss J. P. Cunningham, *Computation-Aware Gaussian Processes: Model Selection And Linear-Time Inference*, Advances in Neural Information Processing Systems (NeurIPS) (2024)

Scaling Insights from Infinite-Width Theory for Next Generation Architectures and Learning Paradigms

Leena Chennuru Vankadara

Scaling is pivotal to the success of modern machine learning. However, this upscaling also introduces new challenges, such as increased training instability. In this talk, I will discuss how infinite-width theory can be utilized to establish optimal scaling rules across various architectures and learning paradigms. I will begin by discussing the scaling behaviour of Multilayer Perceptrons (MLPs) under Sharpness-Aware Minimization—a min-max learning formulation designed to enhance generalization. The analysis extends naturally to other architectures like transformers, ResNets, and CNNs. Additionally, I will discuss the scaling behaviour of structured state space models (SSMs), which have emerged as efficient alternatives to transformers. Owing to the unique structure of their transition matrices, SSMs defy conventional scaling analyses and necessitate specialized approaches. I will discuss the scaling of SSMs within the standard minimization framework, highlighting the need for and implications of specialized scaling strategies.

Learning to act in noisy contexts using deep proxy learning

Arthur Gretton
(joint work with Liyuan Xu, Heishiro Kanagawa)

We consider problem of evaluating the expected outcome of an action or policy, using off-policy observations, where the relevant context is noisy/anonymized. This scenario might arise due to privacy constraints, data bandwidth restrictions, or both. As an example, users might wish to determine the anticipated outcome of an exercise regime, with only an incomplete view available of their fitness levels (for instance, from journaling or wearables). We will employ the recently developed tool of proxy causal learning to address this problem. In brief, two noisy views of the context are used: one prior to the user action, and one subsequent to it, and influenced by the action. This pair of views will allow us to provably recover the average causal effect of an action under reasonable assumptions. As a key benefit of the proxy approach, we need never explicitly model or recover the hidden context. Our implementation employs learned neural net representations for both the action and context, allowing each to be complex and high dimensional (images, text). We demonstrate the deep proxy learning method in a setting where the action is an image, and show that we outperform an autoencoder-based alternative.

References

[1] L. Xu and H. Kanagawa and A. Gretton, *Deep Proxy Causal Learning and its Application to Confounded Bandit Policy Evaluation*, Advances in Neural Information Processing Systems (NeurIPS) 34 (2021).

Beyond the Curse of Dimensionality with Hyper-Kernel Ridge Regression

Shuo Huang

(joint work with Hippolyte Labarrière, Ernesto De Vito, Tomaso Poggio, Lorenzo Rosasco)

Compared to kernel methods, neural networks inherently learn a data representation during the training process. This data-driven representation can adapt to the problem at hand, potentially reducing the need for data and mitigating the curse of dimensionality. The multi-index model (MIM) is a classic statistical framework that can be used to investigate some of these ideas. In this paper, we explore an extension of kernel ridge regression where kernels are composed with a linear representation that can be learned. We refer to this approach as hyper-kernel ridge regression (H-KRR). H-KRR is conceptually similar to neural networks while preserving some of the favorable mathematical properties of kernel methods. Our main contribution is to show that H-KRR can learn MIM effectively from finite samples, thereby escaping the curse of dimensionality. Additionally, some numerical experiments are provided to verify the effectiveness of the algorithm. Our analysis builds on and extends ideas from kernel methods theory to account for the compositional nature of hyper-kernels.

References

[1] S. Huang, H. Labarrière, E. De Vito, T. Poggio, L. Rosasco, *Beyond the Curse of Dimensionality with Hyper-Kernel Ridge Regression*, preprint.

Market Making without Regret

Tom Cesari

(joint work with Nicolò Cesa-Bianchi, Roberto Colomboni, Luigi Foscari, and Vinayak Pathak)

We consider a sequential decision-making setting where, at every round t, a *market maker* posts a *bid* price B_t and an *ask* price A_t to an incoming trader (the *taker*) with a private valuation for one unit of some asset. If the trader's valuation is lower than the bid price, or higher than the ask price, then a trade (sell or buy) occurs. If a trade happens at round t, then letting P_t be the market price (observed only at the end of round t), the maker's utility is $P_t - B_t$ if the maker bought the asset, and $A_t - P_t$ if they sold it. We characterize the maker's regret with respect to the best fixed choice of bid and ask pairs under a variety of assumptions (adversarial, i.i.d., and their variants) on the sequence of market prices and valuations. Our upper bound analysis unveils an intriguing connection relating market making to first-price auctions and dynamic pricing. Our main technical contribution is a lower bound for the i.i.d. case with Lipschitz distributions and independence between prices and valuations. The difficulty in the analysis stems from the unique structure of the reward and feedback functions, allowing an algorithm to acquire information by graduating the "cost of exploration" in an arbitrary way.

References

[1] N. Cesa-Bianchi, T. Cesari, R. Colomboni, L. Foscari, V. Pathak, *Market Making without Regret*, preprint arXiv:2411.13993 (2024).

How to `jit` your `jet`: Accelerating Differential Operators with Linearity

Felix Dangel
(joint work with Marius Zeinhofer)

We explore automating the acceleration of differential operators through compute graph simplifications based on the concept of linearity. These occur in common differential operators like the Laplacian, that computes then sums diagonal elements of the Hessian using Taylor mode automatic differentiation (`jets`). Instead, we show that the Taylor coefficients can first be summed, then propagated, which reduces computational cost. Due to the simplicity of this simplification (propagating a sum up a computation graph), we argue it could (or should) be performed by the just-in-time (`jit`) compiler in machine learning frameworks. Our preliminary experiments achieve promising, fully automated, speed-ups, which we believe can easily be integrated into automatic differentiation libraries.

References

[1] Griewank, A., & Walther, A., *Evaluating derivatives: principles and techniques of algorithmic differentiation*, SIAM (2008).

[2] Li, R., Ye, H., Jiang, D., Wen, X., Wang, C., Li, Z., Li, X., ... *Forward laplacian: a new computational framework for neural network-based variational monte carlo.* (2023).

Rethinking Approximate Gaussian Inference in Classification

Nathaël Da Costa
(joint work with Bálint Mucsányi, Philipp Hennig)

In classification tasks, softmax functions are ubiquitously used as output activations to produce predictive probabilities. Such outputs only capture aleatoric uncertainty. To capture epistemic uncertainty, approximate Gaussian inference methods have been proposed, which output Gaussian distributions over the logit space. Predictives are then obtained as the expectations of the Gaussian distributions pushed forward through the softmax. However, such softmax Gaussian integrals cannot be solved analytically, and Monte Carlo approximations can be costly and noisy. We propose a simple change in the learning objective which allows the *exact* computation of predictives and enjoys improved training dynamics, with no runtime or memory overhead. This framework is compatible with a family of output activation functions that includes the softmax, as well as element-wise normal cumulative distribution function and sigmoid. Moreover, it allows for approximating the Gaussian pushforwards with Dirichlet distributions by analytic moment matching. We evaluate our approach combined with several approximate Gaussian inference methods (Laplace, HET, SNGP) on large- and small-scale datasets (ImageNet, CIFAR-10), demonstrating improved uncertainty quantification capabilities compared to softmax models with Monte Carlo sampling.

References

[1] B. Mucsányi, N. Da Costa, P. Hennig, *Rethinking Approximate Gaussian Inference in Classification*, arXiv:2502.03366 (2025).

Empirical risk minimization for risk-neutral composite optimal control with applications to bang-bang control

Daniel Walter
(joint work with Johannes Milz)

Nonconvex optimization problems governed by differential equations arise in a multitude of application areas, such as sensor placement, resource assessment of renewable tidal energy, and design of groundwater remediation systems. In this talk, we consider risk-neutral composite optimal control problems where the objective function is the sum of a potentially nonconvex expectation function and a nonsmooth convex function. While a proper choice of the latter promotes favourable structural features of the obtained minimizers, it also significantly complicates, both, the theoretical analysis of the problem as well as its realization, e.g. due to a lack of strong convexity. In the present setting, a further layer of difficulty is added by random parameters in the underlying equation. To approximate the risk-neutral optimization problems, we use a Monte Carlo sample-based approach, study its asymptotic consistency, and derive nonasymptotic sample size estimates relying on a covering number approach. Our analyses leverage problem structure commonly encountered in PDE-constrained optimization problems, including compact embeddings and growth conditions. We apply our findings to bang-bang-type optimal control problems and propose the use of a conditional gradient method to solve them effectively. Numerical illustrations are presented for, both, linear as well as bilinear elliptic PDEs demonstrating the sharpness of some of the obtained results.

References

[1] J. Milz, D. Walter *Empirical risk minimization for risk-neutral composite optimal control with applications to bang-bang control*, `https://arxiv.org/abs/2408.10384`.

Bandits for free in Multiclass Classification

Yishay Mansour

(joint work with Liad Erez, Alon Cohen, Tomer Koren, Shay Moran)

We revisit the classical problem of multiclass classification with bandit feedback (Kakade, Shalev-Shwartz and Tewari, 2008), where each input classifies to one of K possible labels and feedback is restricted to whether the predicted label is correct or not. Our primary inquiry is with regard to the dependency on the number of classes K, which is often very large in multiclass problems. I will survey two recent results in this context:

(1) A characterization of the minimax regret of bandit multiclass, establishing that it is of the for $\min\{|H| + \sqrt{T}, \sqrt{KT \log(|H|)}\}$ for finite hypothesis class H; In particular, we present a new bandit classification algorithm that guarantees this rate, improving over classical algorithms (such as EXP4) for moderately-sized hypothesis classes, and give a matching lower bound establishing tightness (up to log-factors) in all parameter regimes.

(2) A novel learning algorithm for the agnostic PAC version of the problem, with sample complexity of $O((poly(K) + 1/\epsilon^2) \log(|H|/\delta))$; our algorithm utilizes a stochastic optimization technique to minimize a log-barrier potential based on Frank-Wolfe updates for computing a low-variance exploration distribution over the hypotheses, and is made computationally efficient provided access to an ERM oracle over H.

We also provide an extension general classes and establish similar sample complexity bounds in which $\log |H|$ is replaced by the Natarajan dimension. Surprisingly, these results match the asymptotic optimal rates with full-information, and reveal a stark contrast between the PAC and regret-minimization versions of the problem.

References

[1] L Erez, A Cohen, T Koren, Y Mansour, S Moran, *Fast Rates for Bandit PAC Multiclass Classification*, NeurIPS 2024

[2] L Erez, A Cohen, T Koren, Y Mansour, S Moran , *The Real Price of Bandit Information in Multiclass Classification* , COLT 2024.

Denoising diffusion models without diffusion

Francis Bach

(joint work with Saeed Saremi, Ji-wei Liao)

Denoising diffusion models lead to state-of-the-art performance for sampling complex objects. They are based on the idea of sampling by denoising from Hyvärinen (2005), the possibility of learning denoiser from empirical data (the key idea of Vincent), by score matching, and the possibility of progressive denoising. The progressive denoising can be achieved through continuous-time stochastic processes (diffusions), as done by Song & Ermon (2019).

In this work, we propose a diffusion-free view that can be readily extended to discrete data.

References

[1] A. Hyvärinen, *Estimation of Non-Normalized Statistical Models by Score Matching*, Journal of Machine Learning Research (2005).

[2] Y. Song, S. Ermon *Generative Modeling by Estimating Gradients of the Data Distribution*, NeurIPS (2019).

Attention-based predictors and single location regression

Claire Boyer

(joint work with Pierre Marion, Raphaël Berthier, Gérard Biau)

Attention-based models, such as Transformer, excel across various tasks but lack a comprehensive theoretical understanding, especially regarding token-wise sparsity and internal linear representations. To address this gap, we introduce the single-location regression task, where only one token in a sequence determines the output, and its position is a latent random variable, retrievable via a linear projection of the input. To solve this task, we propose a dedicated predictor, which turns out to be a simplified version of a non-linear self-attention layer. We study its theoretical properties, by showing its asymptotic Bayes optimality and analyzing its training dynamics. In particular, despite the non-convex nature of the problem, the predictor effectively learns the underlying structure. This work highlights the capacity of attention mechanisms to handle sparse token information and internal linear structures.

References

[1] P. Marion, R. Berthier, G. Biau, & C. Boyer. (2024). *Attention layers provably solve single-location regression*, Proceedings of the International Conference on Learning Representations (ICLR 2025). arXiv preprint arXiv:2410.01537.

Causal Representation Learning with the Invariance Principle

Francesco Locatello

(joint work with Dingling Yao, Dario Rancati, Riccardo Cadei, Marco Fumero)

Machine learning and AI have the potential to transform data-driven scientific discovery, enabling not only accurate predictions for several scientific phenomena but also causal understanding. Toward this, we present a new framework for causal representation learning based on the invariance principle that generalizes most existing methodologies across levels of the causal hierarchy [1]. Thanks to the increased flexibility, we show improved performance on our ISTAnt data set, the first real-world benchmark for estimating causal effects from high-dimensional observations in experimental ecology [3]. Further, we connect causal representation learning with recent advances in dynamical systems discovery that, when combined, enable learning scalable and controllable models with identifiable trajectory-specific parameters [2], which we apply to real-world climate data.

References

[1] Yao, D., et al.*Unifying Causal Representation Learning with the Invariance Principle*, to appear at International Conference on Learning Representations, ICLR 2025.

[2] Yao, D., et al.*Marrying Causal Representation Learning with Dynamical Systems for Science*, Advances in Neural Information Processing Systems, NeurIPS 2024.

[3] Cadei, R., et al.*Smoke and Mirrors in Causal Downstream Tasks*, Advances in Neural Information Processing Systems, NeurIPS 2024.

Mind the Spikes: Benign Overfitting of Kernels and Neural Networks in Fixed Dimension

Moritz Haas

(joint work with David Holzmüller, Ulrike von Luxburg, Ingo Steinwart)

The success of over-parameterized neural networks trained to near-zero training error has caused great interest in the phenomenon of benign overfitting, where estimators are statistically consistent even though they interpolate noisy training data. While benign overfitting in fixed dimension has been established for some learning methods, most of the literature suggests that for regression with typical kernel methods and wide neural networks, benign overfitting requires a high-dimensional setting where the dimension grows with the sample size. In this talk, we show that the smoothness of the estimators, and not the dimension, is the key: benign overfitting is possible if and only if the estimator's derivatives are large enough. We generalize existing inconsistency results to non-interpolating models and more kernels to show that benign overfitting with moderate derivatives is impossible in fixed dimension. Conversely, we show that rate-optimal benign overfitting is possible for regression with a sequence of spiky-smooth kernels with large derivatives. Using neural tangent kernels, we translate our results to wide neural networks. We prove that while infinite-width networks do not overfit benignly with the ReLU activation, this can be fixed by adding small high-frequency fluctuations to the activation function. Our experiments verify that such neural networks, while overfitting, can indeed generalize well even on low-dimensional data sets.

References

[1] M. Haas, D. Holzmüller, U. von Luxburg, I. Steinwart. *Mind the spikes: Benign overfitting of kernels and neural networks in fixed dimension.* Advances in Neural Information Processing Systems, NeurIPS 2023.

LUNO: Linearization turns neural operators into function-valued GPs

Emilia Magnani

(joint work with Marvin Pförtner, Tobias Weber, Philipp Hennig)

Neural operators generalize neural networks to learn mappings between function spaces from data. They are commonly used to learn solution operators of parametric partial differential equations (PDEs) or propagators of time-dependent PDEs. However, to make them useful in high-stakes simulation scenarios, their inherent predictive error must be quantified reliably. We introduce LUNO, a novel framework for approximate Bayesian uncertainty quantification in trained neural operators. Our approach leverages model linearization to push (Gaussian) weight-space uncertainty forward to the neural operator's predictions. We show that this can be interpreted as a probabilistic version of the concept of currying from functional programming, yielding a function-valued (Gaussian) random process belief. Our framework provides a practical yet theoretically sound way to apply existing Bayesian deep learning methods such as the linearized Laplace approximation to neural operators. Just as the underlying neural operator, our approach is resolution-agnostic by design. The method adds minimal prediction overhead, can be applied post-hoc without retraining the network, and scales to large models and datasets. We evaluate these aspects in a case study on Fourier neural operators.

References

[1] E. Magnani, M. Pförtner, T. Weber, and P. Hennig, *Linearization Turns Neural Operators into Function-Valued Gaussian Processes*, arXiv preprint arXiv:2406.05072, 2024.

Associative Memories as a Building Block in Transformers

Alberto Bietti

(joint work with Vivien Cabannes, Elvis Dohmatob, Diane Bouchacourt, Hervé Jegou, Leon Bottou, Eshan Nichani, Jason Lee)

Large language models based on transformers have achieved great empirical success. However, as they are deployed more widely, there is a growing need to better understand their internal mechanisms in order to make them more reliable. These models appear to store vast amounts of knowledge from their training data, and to adapt quickly to new information provided in their context or prompt. Through toy tasks for reasoning and factual recall, we highlight the role of weight matrices as associative memories, and provide theoretical results on how gradients enable their learning during training, as well as how over-parameterization affects their storage capacity. Specifically, the associative memory building block takes the following form

$$W = \sum_{z \in [N]} u_{f(z)} e_z^\top \in \mathbb{R}^{d \times d},$$

where d is the model width, $e_z, u_y \in \mathbb{R}^d$ are random input/output embeddings, and N is the number of target associations $(z, f(z))$ that we want to store. This structure arises naturally from gradients at intermediate layers of deep networks, and is amenable to a precise study of memorization capacity at finite width.

References

[1] A. Bietti, V. Cabannes, D. Bouchacourt, H. Jegou, L. Bottou *Birth of a Transformer: A Memory Viewpoint*, NeurIPS (2023).

[2] V. Cabannes, E. Dohmatob, A. Bietti. *Scaling Laws for Associative Memories*, ICLR (2024).

[3] E. Nichani, J.D. Lee, A. Bietti. *Understanding Factual Recall in Transformers via Associative Memories*, ICLR (2025).

Nonsmooth optimization for machine learning

Andrea Walther

(joint work with Aswin Kannan, Timo Kreimeier)

The retail industry is governed by crucial decisions on inventory management, discount offers and stock clearings yielding various optimization problems. One important task is to learn the demand (sales) elasticity with respect to product prices. Another one is the dynamic revenue maximization problem, which takes in the coefficients of demand as inputs. While both tasks present nonsmooth and data-driven optimization problems, the latter is a challenging nonlinear problem in massive dimensions that is also subject to constraints. Traditional approaches to learn the corresponding parameters relied on using reformulations and approximations, thereby leading to potentially suboptimal solutions. In this work, we exploit the nonsmooth structure generated by the piecewise linear and piecewise smooth structure of the target function. Further details can be found in [1]. Furthermore, we adapted the Constrained Active Signature Method (CASM) [2] to solve the resulting tasks. Two real world retail examples (UK and US market data from 2017–2019) and one simulated use-case are studied from an empirical standpoint. Numerical results demonstrate good performance of our approach.

References

[1] A. Kannan, T. Kreimeier and A. Walther: On solving nonsmooth retail portfolio maximization problem using active signature methods. TRR 154 preprint 520.

[2] T. Kreimeier, A. Walther and A. Griewank: Constrained Piecewise Linear Optimization by an Active Signature Method. In: Mathematical Optimization for Machine Learning (Proceedings of the MATH+ TES summer semester 2023), de Gruyter, to appear in 2025.

Probabilistic Functional Programming

Marvin Pförtner

(joint work with Jonathan Wenger, Nathaël Da Costa, Lancelot Da Costa, Tim Weiland, Emilia Magnani, Tobias Weber, Jon Cockayne, Ingo Steinwart, Philipp Hennig)

Most approaches to supervised machine learning aim to infer an unknown function from pointwise observations. However, especially in scientific inference, we often encounter more general sources of information about an unknown function such as the fact that a partial differential equation (PDE) holds. In this talk, I will argue that a principled approach to nonparametric Bayesian inference under such information operators requires functions to be treated as first-class citizens, just like in functional programming. I will give an introduction to the theory of Gaussian measures on (infinite-dimensional) real vector spaces and demonstrate how this framework can be used to make Gaussian process inference with observations made through linear functionals rigorous [1, 2]. Finally, I will show two applications of this theoretical result: probabilistic numerical PDE solvers [1, 3, 4], and a method for approximate Bayesian uncertainty quantification in neural operators that is based on function-valued Gaussian processes [6, 7].

References

[1] M. Pförtner, I. Steinwart, P. Hennig, and J. Wenger, *Physics-Informed Gaussian Process Regression Generalizes Linear PDE Solvers*, arXiv preprint arXiv:2212.12474, 2022.

[2] N. Da Costa, M. Pförtner, L. Da Costa, and P. Hennig, *Sample Path Regularity of Gaussian Processes from the Covariance Kernel*, arXiv preprint arXiv:2312.14886, 2023.

[3] T. Weiland, M. Pförtner, and P. Hennig, *Scaling up Probabilistic PDE Simulators with Structured Volumetric Information*, arXiv preprint arXiv:2406.05020, 2024.

[4] T. Weiland, M. Pförtner, and P. Hennig, *Flexible and Efficient Probabilistic PDE Solvers through Gaussian Markov Random Fields*, 2025.

[5] M. Pförtner, J. Wenger, J. Cockayne, and P. Hennig, *Computation-Aware Kalman Filtering and Smoothing*, arXiv preprint arXiv:2405.08971, 2024.

[6] E. Magnani, M. Pförtner, T. Weber, and P. Hennig, *Linearization Turns Neural Operators into Function-Valued Gaussian Processes*, arXiv preprint arXiv:2406.05072, 2024.

[7] T. Weber, E. Magnani, M. Pförtner, and P. Hennig, *Uncertainty Quantification for Fourier Neural Operators.* In: ICLR 2024 Workshop on AI4DifferentialEquations In Science, 2024.

Causal de Finetti & Do Finetti: On causality for exchangeable data

Siyuan Guo

Like many machine learning methods, causality is developed based on the assumption of independent and identically distributed (i.i.d.) data. However, it is well-known that even with infinite i.i.d. data, constraint-based causal discovery methods can only identify causal structures up to broad Markov equivalence classes, posing fundamental limitations for causal discovery.

In causal de Finetti, we observe that exchangeable data contains richer conditional independence structure than i.i.d. data. This richer structure can be leveraged for causal discovery. Do Finetti further builds on the causal framework and studies interventions by establishing formal do-calculus in exchangeable data and proving generalized truncated factorization for identification and computation of causal effects.

Algorithmic Collective Action in Machine Learning

Celestine Mendler-Dünner

(joint work with Moritz Hardt, Eric Mazumdar, Tijana Zrnic)

I present a simple theoretical model to initiate a principled study of algorithmic collective action in learning systems [1]. It describes a collective interacting with a firm that deploys a machine learning algorithm. Each individual in the collective controls a single training data point and together they execute an algorithmic strategy to modify their data and achieve a collective goal. In three fundamental learning-theoretic settings – the case of a nonparametric optimal learning algorithm, a parametric risk minimizer, and gradient-based optimization – I present coordinated algorithmic strategies and characterize natural success criteria as a function of the collective's size. Complementing our theory, I present systematic experiments on a skill classification task using a transformer-based language model and demonstrate a striking correspondence between our empirical observations and the predictions made by our analysis. Taken together, our results broadly support the conclusion that algorithmic collectives of exceedingly small fractional size can exert significant control over a platform's learning algorithm.

References

[1] M. Hardt, E. Mazumdar, C. Mendler-Dünner, and T. Zrnic, *Algorithmic collective action in machine learning*, Proceedings of the International Conference on Machine Learning (2023).

Optimal transport distances for Markov chains

Gergely Neu

(joint work with Sergio Calo, Anders Jonsson, Ludovic Schwartz,
and Javier Segovia-Aguas)

How can one define similarity metrics between stochastic processes? Understanding this question can help us design better representations for dynamical systems, study distances between structured objects, formally verify complex programs, and so on. In the past, the dominant framework for studying this question has been that of bisimulation metrics, a concept coming from theoretical computer science. My recent work has been exploring an alternative perspective based on the theory of optimal transport, which has led to surprising results, including a proof of the fact that bisimulation metrics are, in fact, optimal transport distances. This realization allowed us to import tools from optimal transport and develop computationally efficient methods for computing distances between Markov chains via the reduction of the problem to a finite-dimensional linear program. In this talk, I have introduced this framework and the foundations of the most recent algorithmic developments, as well as discussed the potential for representation learning in more detail.

Are self-supervised models doing kernel PCA?

Debarghya Ghoshdastidar

(joint work with Gautham G. Anil, Pascal Esser, Maximilian Fleissner)

This talk will have two parts. In the first part, I will briefly introduce self-supervised pretraining commonly used in foundation models for vision and tabular data. I will also introduce the key question related to statistical generalisation in foundation models: *How do we guarantee statistical generalisation for different downstream prediction tasks given that the model is pre-trained with large amount of unlabelled (augmented) data?* I will conclude this part with some initial ideas and results assuming that the foundation model uses a kernel-based encoder to learn representations via self-supervised pre-training.

In the second part part, I will focus on the equivalence between self-supervised neural networks and kernel principal component analysis (PCA). This equivalence is based on two ideas: (i) optimal solution of self-supervised kernel models can be computed as a spectral embedding [1], and (ii) infinitely wide neural networks are equivalent to kernel models, characterised by the neural tangent kernel (NTK). I conclude with our recent results on the convergence to the NTK under self-supervised training [2, 3].

References

[1] P. Esser, M. Fleissner, and D. Ghoshdastidar, *Non-parametric representation learning with kernels* In:Proceedings of the AAAI Conference on Artificial Intelligence **38** (2024), 11910-11918. arXiv preprint arXiv:2309.02028.

[2] G. G. Anil, P. Esser, and D. Ghoshdastidar, *When can we approximate wide contrastive models with neural tangent kernels and principal component analysis?* In:Proceedings of the AAAI Conference on Artificial Intelligence **39** (2025). arXiv preprint arXiv:2403.08673.

[3] M. Fleissner, G. G. Anil, and D. Ghoshdastidar, *Infinite width limits of self supervised neural networks* In:Proceedings of The 28th International Conference on Artificial Intelligence and Statistics, PMLR (2025). arXiv preprint arXiv:2411.11176.

Achievable distributional robustness when the robust risk is only partially identified

Fanny Yang

(joint work with Julia Kostin, Nicola Gnecco)

In safety-critical applications, machine learning models should generalize well under worst-case distribution shifts, that is, have a small robust risk. Invariance-based algorithms can provably take advantage of structural assumptions on the shifts when the training distributions are heterogeneous enough to identify the robust risk. However, in practice, such identifiability conditions are rarely satisfied –

a scenario so far underexplored in the theoretical literature. In this paper, we aim to fill the gap and propose to study the more general setting when the robust risk is only partially identifiable. In particular, we introduce the worst-case robust risk as a new measure of robustness that is always well-defined in this setting. Its minimum corresponds to an algorithm-independent (population) minimax quantity that measures the best achievable robustness under partial identifiability. While these concepts can be defined more broadly, in this paper we introduce and derive them explicitly for a linear model for concreteness of the presentation. Specifically, we prove how previous approaches rank differently in terms of worst-case robust risk and are suboptimal in the partially identifiable case. We then evaluate these approaches and the minimizer of the (empirical) worst- case robust risk on and find a similar trend: the test error of existing robustness methods grows increasingly suboptimal as the fraction of data from unseen environments increases.

References

[1] Julia Kostin and Nicola Gnecco and Fanny Yang, *Achievable distributional robustness when the robust risk is only partially identified*, Proceedings of the Annual Conference on Neural Information Processing Systems (2024).

Why does Adam work so well? Heavy-tailed class imbalance in language models

Frederik Künstner

(joint work with Alberto Bietti, Jacques Chen, Wilder Lavington, Alan Milligan, Mark Schmidt, and Robin Yadav)

Adam has been shown to outperform gradient descent in optimizing large language transformers empirically, and by a larger margin than on other tasks, but it is unclear why this happens. We show that the heavy-tailed class imbalance found in language modeling tasks leads to difficulties in the optimization dynamics. When training with gradient descent, the loss associated with infrequent words decreases slower than the loss associated with frequent ones. As most samples come from relatively infrequent words, the average loss decreases slowly with gradient descent. On the other hand, Adam and sign-based methods do not suffer from this problem and improve predictions on all classes. To establish that this behavior is indeed caused by class imbalance, we show empirically that it persist through different architectures and data types, on language transformers, vision CNNs, and linear models. We study this phenomenon on a linear classification with cross-entropy loss, showing that heavy-tailed class imbalance leads to ill-conditioning, and that the normalization used by Adam can counteract it.

References

[1] Frederik Kunstner, Alan Milligan, Robin Yadav, Mark Schmidt, and Alberto Bietti. *Heavy-tailed class imbalance and why adam outperforms gradient descent on language models* Advances in Neural Information Processing, 2024.

[2] Frederik Kunstner, Jacques Chen, Jonathan Wilder Lavington, and Mark Schmidt. *Noise is not the main factor behind the gap between sgd and adam on transformers, but sign descent might be*, International Conference on Learning Representations, 2023.

Towards Practical Probabilistic PDE Solvers

Tim Weiland

(joint work with Marvin Pförtner, Philipp Hennig)

Physical simulations are subject to various sources of uncertainty arising from noisy or missing measurements, unknown parameters, and discretization error of a numerical PDE solver. Probabilistic PDE solvers generalize classic numerical solvers while providing a principled treatment of these uncertainties. Yet, despite their theoretical properties, such solvers have not seen much practical use so far. In my talk, I discuss two techniques to greatly improve the scalability of probabilistic PDE solvers. First, viewing a PDE as an "infinite data source" in the sense of probabilistic numerics, we consider the idea of alternative ways of "drawing information" from this data source. Concretely, we construct volumetric information operators which reproduce the finite volume method, and discuss efficient computational techniques. Second, we challenge the popular approach to probabilistic PDE solvers of using a Gaussian process prior in covariance function form. We see that stochastic PDEs allow for a much more rich, physically meaningful expression of prior knowledge, while simultaneously providing drastic computational benefits through Gaussian Markov Random Field priors which allow for the use of highly efficient sparse linear algebra.

References

[1] T. Weiland, M. Pförtner, and P. Hennig, *Scaling up Probabilistic PDE Simulators with Structured Volumetric Information*, arXiv preprint arXiv:2406.05020, 2024.

[2] T. Weiland, M. Pförtner, and P. Hennig, *Flexible and Efficient Probabilistic PDE Solvers through Gaussian Markov Random Fields*, 2025.

What is a causal representation?

Bernhard Schölkopf

Epistemology is traditionally concerned with human knowledge, yet it extends to artificially intelligent agents. We aim to contribute to its formalization by analyzing processes of model building in artificial intelligence (AI). Given that our access to the world is an indirect one, mediated by processes of sensory transduction, can we nevertheless build models that are structurally related to the world?

For an active agent, it is not sufficient for a world model to represent what is there, as done in the popular paradigm of statistical representation learning. Representations should also support interventions, i.e., include causal information. Ultimately, interventional representations could form the basis of *thinking* as *acting in an imagined space* [4]. A key obstacle preventing the use of causal models in

AI, however, is the requirement that the relevant causal variables or “symbols” be specified a priori and/or directly observed. This underscores the need to combine causal modeling with representation learning. What makes a good representation?

Leibniz developed the following thought [2]. Suppose that the concept of ‘living being’ were represented by the number 2, and the concept ‘rational’ were represented by the number 3. Then the concept of ‘human’ would correspond to the product of 2 and 3, yielding 6. This is an example of a homomorphism, where structure in one domain (here, logical relationship) is expressed in another domain (integer multiplication). By expressing it in another space, homomorphisms preserve relevant structure while making it easier to handle.

Helmholtz asked himself how we arrive at reliable perceptions. Hypothesizing that we make predictions and test them sensori-motorically, he viewed perceptions as unconscious inductive inferences. Pragmatic criteria for testing representations may involve actions and their perceptual consequences. Hertz argued that *we create internal illusions or symbols of external objects, and we do so in such a way that the images’ consequences that are necessary in thought always coincide with the images of the depicted objects’ consequences that are necessary in nature* [1]. This form of consistency [6] constitutes a desideratum for causal representations.

To gain further intuition on the link between the world and models thereof, let us consider an example [7]: if variations of natural lighting (the position of the sun, clouds, etc.) imply that the visual environment can appear in brightness conditions spanning orders of magnitude, then visual processing algorithms in our nervous system should employ methods that represent (and thus factor out) these variations, rather than building separate sets of face recognizers, say, for every lighting condition. If our brain were to compensate for the lighting changes by a gain control mechanism, say, then this mechanism in itself need not have anything to do with the physical mechanisms bringing about brightness differences. It would, however, play a role in a world model’s mathematical structure akin to the role the physical mechanisms play in the world’s modular structure.

Causal representations should preserve not just statistical, but also interventional information. Consider a class of maps $A : \mathcal{X} \to \mathcal{X}$, each representing the physical operation of an intervention on the world, observed through $\mathbf{X} \in \mathcal{X}$. A **causal representation** comprises a map ϕ taking an observation $\mathbf{X}$ to $\mathbf{Z} \in \mathcal{Z}$, and a map Φ taking an intervention to a map $\mathcal{Z} \to \mathcal{Z}$, such that (ϕ, Φ) preserve interventional information in the sense that the effect of interventions can consistently be computed in the space $\mathcal{Z}$, i.e., the below diagram commutes.

$$\begin{tikzcd}
\mathcal{Z} \arrow[r, "\Phi(A)"] & \mathcal{Z} \\
\mathcal{X} \arrow[u, "\phi"] \arrow[r, "A"] & \mathcal{X} \arrow[u, "\phi"]
\end{tikzcd}$$

In a biological system, ϕ might be a sensory mapping implemented by the eye and the visual cortex, while Φ might be realized through *efference copies*, i.e., internal copies of motor signals generated by an animal’s brain. In order to realize an internal world model, those copies should then be processed and represented in $\mathcal{Z}$ where they can be related to sensory representations generated by $\boldsymbol{\phi}$.

While the above implements Hertz' notion of consistency, it also allows for incorporation of homomorphism. E.g., if we assume a group structure on the class of interventions, we may require that Φ be a homomorphism to a group of linear operators on a suitable $\mathcal{Z}$, i.e., $\Phi(A_1) \circ \Phi(A_2) = \Phi(A_1 \circ A_2)$ [3].

Imagine, for instance, that we have a set of objects making up a scene. Assume that $\mathcal{X}$ contains high-dimensional images of the scene, whereas $\mathcal{Z}$ only contains location coordinates of the objects. Consider an intervention that shifts the position of one object along one coordinate: this intervention is complicated in $\mathcal{X}$, (nonlinearly) affecting many pixels at the same time, yet it corresponds to a simple change to one coordinate in $\mathcal{Z}$. Only representing the positions of objects is not a rich representation; however, if that is all we can intervene upon, then there is no need for more. If other interventions are possible, e.g., changing the illumination, then $\boldsymbol{\phi}$ should represent additional parameters, such as appearance or material properties of the objects in the scene.

In causal modeling, we are often dealing with probability distributions over the observations $\mathbf{X}$ and the latent variables $\mathbf{Z}$ respectively. Recall that $A(\mathbf{X})$ is the result of applying the intervention A on $\mathbf{X}$, and assume that in the latent space $\mathcal{Z}$, we have an interventional calculus *do* [5] that allows us to compute the distribution $P_{\mathbf{Z}}^{do(A)}$ obtained by performing an intervention A. Then the commutative diagram can be expressed as

$$(\boldsymbol{\phi}_*(P_{\mathbf{X}}))^{do(\Phi(A))} = \boldsymbol{\phi}_*(A_*(P_{\mathbf{X}})), \tag{1}$$

where the asterisk $*$ denotes the push-forward of probability distributions. The mapping Φ on A needs to be known or estimated from data. Multiple mappings may satisfy the condition, differing in which information they capture and how well they disentangle the interventions. Let us consider several special cases.

Statistical representations. If the class of interventions is trivial (i.e., only containing the identity), (1) is satisfied irrespective of $\boldsymbol{\phi}$.

Causal graphs. Causal graphs are directed acyclic graphs $\mathcal{G}$ where each node represents a variable and comes with a "mechanism" modeled as a conditional of the node given its parents, and the "do-calculus" [5] tells us how to modify the entailed joint distribution when interventions (i.e., changes to some conditionals) occur. In $(\boldsymbol{\phi}_*(P_{\mathbf{X}}))^{do(\Phi(A))}$, the term $\boldsymbol{\phi}_*(P_{\mathbf{X}})$ is the observational distribution mapped into the latent space, and the notation $do(\Phi(A))$ denotes that we are applying the rules of the do-calculus with respect to $\mathcal{G}$ to update the distribution $\boldsymbol{\phi}_*(P_{\mathbf{X}})$ according to the intervention A. We thus think of $\mathcal{G}$ as part of (the image under) Φ. Learning Φ includes learning a causal graph $\mathcal{G}$ along with its mechanisms and the intervention targets such that the given interventions A can be realized as do-interventions in the graph.

Invertible representations. A class of representations that satisfy (1) are those where $\boldsymbol{\phi}$ is invertible, and $\Phi(A) := \boldsymbol{\phi} \circ A \circ \boldsymbol{\phi}^{-1}$. One such case is Independent

Component Analysis (ICA), which can be used for source separation. Imagine we have several independent acoustic sources and we observe a mixture $\mathbf{X}$, recorded by several microphones. Assume that our interventions A affect the volume of one source at a time. Observed through $\mathbf{X}$, each intervention simultaneously changes the volume in all microphones, by different amounts, giving us a vector each. Putting the vector obtained for the different source interventions next to each other, we get a matrix A that characterizes how the interventions affect our observations $\mathbf{X}$. ICA aims to find a matrix $\boldsymbol{\phi}$ diagonalizing the intervention matrix. The action of our interventions mapped in the representation space, i.e., $\Phi(A) := \boldsymbol{\phi} \circ A \circ \boldsymbol{\phi}^{-1}$, will correspond to atomic intervention affecting one $\mathbf{Z}_i$ at a time. In practice, subject to suitable conditions, ICA is carried out by computing $\boldsymbol{\phi}$ such as to minimize the dependence between the variables $\mathbf{Z}_i$. In this case, the $\mathbf{Z}_i$ correspond (up to certain degrees of freedom) to the sources. One can think of ICA as the process of fitting a simple causal model to the observed data [8].

Generative setting. A fruitful point of view, further generalizing that of ICA, can be developed for the case where we assume that the observations $\mathbf{X}$ are the result of applying a *generative* mapping ψ to unobserved *physical* variables $\mathbf{Z}$.

Disentangled causal representations. In practice, we are interested in mappings that leads to interventions acting sparsely yet nontrivially. E.g., if all $\Phi(A)$ only affect single mechanisms on $\mathcal{Z}$, then the resulting push forward model recovers what has been called the *causal disentangled factorization*, with factors satisfying the Independent Causal Mechanisms condition [7]. We may call a causal representation $(\boldsymbol{\phi}, \Phi)$ **disentangled** w.r.t. a class of interventions A if every $\Phi(A)$ is an atomic intervention, i.e., it affects only a single variable (or causal mechanism). We may call it **faithful** with respect to a set of interventions if $\Phi : A \mapsto \Phi(A)$ is injective.

Note, finally, that our notion of a causal representation does not require the existence of a ground truth generative model for the data. Rather, it is *instrumental*: the value of a causal representation is not determined by whether it recovers a hypothetical ground truth, but by whether it allows us to consistently compute the effect of actions or interventions.

References

[1] H. Hertz, *Prinzipien der Mechanik. In neuem Zusammenhange dargestellt.* Johann Ambrosius Barth, Leipzig, 1894.

[2] J. Jost. *Leibniz und die moderne Naturwissenschaft.* Springer, Berlin, Heidelberg, 2019.

[3] H. Keurti, H.-R. Pan, M. Besserve, B. Grewe, and B. Schölkopf. *Homomorphism AutoEncoder – learning group structured representations from observed transitions.* Proceedings of the 40th International Conference on Machine Learning, pages 16190–16215. PMLR, 2023.

[4] K. Lorenz. *Die Rückseite des Spiegels.* R. Piper & Co. Verlag, 1973.

[5] J. Pearl. *Causality: Models, Reasoning, and Inference.* Cambridge University Press, 2009.

[6] P. Rubenstein, S. Weichwald, S. Bongers, J. Mooij, D. Janzing, M. Grosse-Wentrup, and B. Schölkopf. *Causal consistency of structural equation models.* 33rd Conference on Uncertainty in Artificial Intelligence, pages 808–817. Curran Associates, Inc., 2017.

[7] B. Schölkopf. *Causality for machine learning.* Probabilistic and Causal Inference: The Works of Judea Pearl, pages 765–804. Association for Computing Machinery, 2022. arXiv:1911.10500.

[8] L. Wendong, A. Kekić, J. von Kügelgen, S. Buchholz, M. Besserve, L. Gresele, and B. Schölkopf. *Causal component analysis.* Advances in Neural Information Processing Systems, volume 36, 2023.

Reporter: Nathaël Da Costa

Participants

Dr. Francis Bach
INRIA
Departement d'Informatique
Ecole Normale Superieure
Voie DQ 12
2, rue Simone Iff
75012 Paris Cedex
FRANCE

Robi Bhattacharjee
Department of Computer Science
Universität Tübingen
72074 Tübingen
GERMANY

Dr. Alberto Bietti
Flatiron Institute
Simons Foundation
162 5th Avenue
New York, NY 10010
UNITED STATES

Dr. Claire Boyer
Laboratoire de Probabilités, Statistique
et Modélisation (LPSM), BP 158
Sorbonne Université
Campus Pierre et Marie Curie
4, place Jussieu
75252 Paris Cedex 05
FRANCE

Prof. Dr. Nicolò Cesa-Bianchi
Dept. of Computer Science
Università degli Studi di Milano
via Celoria 18
20133 Milano
ITALY

Tom Cesari
School of Electrical Engineering and
Computer Science
Department of Mathematics & Statistics
University of Ottawa
800 King Edward Ave
Ottawa, ON K1N 6N5
CANADA

Dr. Leena Chennuru Vankadara
Amazon Research
Friedrich-Miescher-Straße 4
72076 Tübingen
GERMANY

Nathaël Da Costa
University of Tübingen
72074 Tübingen
GERMANY

Felix Dangel
Vector Institute
Schwartz Reisman Innovation Campus
W1140-108 College Street
Toronto M5G 0C6
CANADA

Dr. Emmanuel Esposito
Department of Computer Science
Università degli Studi di Milano
Via Celoria, 18
20133 Milano
ITALY

Prof. Dr. Nicolas Flammarion
TML
EPFL IC
INJ 336
Station 14
1015 Lausanne
SWITZERLAND

Prof. Dr. Debarghya Ghoshdastidar
TUM School of Computation,
Information and Technology
Technische Universität München
Boltzmannstr. 3
85748 München
GERMANY

Prof. Dr. Arthur Gretton
Gatsby Computational Neuroscience
Unit
25 Howland Street
W1T 4JG London
UNITED KINGDOM

Dr. Siyuan Guo
Max Planck Institute for Intelligent
Systems
Max-Planck-Ring 4
72076 Tübingen
GERMANY

Moritz Haas
Department of Computer Science
Universität Tübingen
Maria-von-Linden-Str. 6
72076 Tübingen
GERMANY

Prof. Dr. Philipp Hennig
Universität Tübingen
Fachbereich Informatik
Lehrstuhl für die Methoden des
Maschinellen Lernens
Maria-von-Linden-Straße 6
72076 Tübingen
GERMANY

Dr. Shuo Huang
IIT Genova
Via Morego 30
16163 Genova
ITALY

Julia Kostin
Institut für Informatik
ETH-Zürich
8092 Zürich
SWITZERLAND

Prof. Dr. Andreas Krause
ETH Zürich
Institute for Machine Learning
OAT Y13.1
Andreasstrasse 5
8092 Zürich
SWITZERLAND

Dr. Frederik Künstner
INRIA
75013 Paris
FRANCE

Dr. Gil Kur
ETH Zürich
8400 Zürich
SWITZERLAND

Dr. Francesco Locatello
Institute of Science and Technology
Austria (IST Austria)
Am Campus 1
3400 Klosterneuburg
AUSTRIA

Emilia Magnani
Universität Tübingen
Fachbereich Informatik
Maria-von-Linden-Straße 6
72076 Tübingen
GERMANY

Prof. Dr. Yishay Mansour
School of Computer Science
Tel Aviv University
69978 Ramat Aviv, Tel Aviv
ISRAEL

Dr. Celestine Mendler-Dünner
ELLIS Institute Tübingen
Maria-von-Linden-Straße 2
& Max Planck Institute
for Intelligent Systems
Max-Planck-Ring 4
72076 Tübingen
GERMANY

Prof. Dr. Nicole Mücke
Institut für Mathematische
Stochastik der TU Braunschweig
Postfach 3329
38023 Braunschweig
GERMANY

Gergely Neu
Universitat Pompeu Fabra
C/ Roc Boronat 138
08018 Barcelona
SPAIN

Dr. Antonio Orvieto
Max Planck Institute for Intelligent
Systems
Max-Planck-Ring 4
72076 Tübingen
GERMANY

Prof. Dr. Vianney Perchet
Crest, ENSAE
Criteo AI Lab
5 Avenue Le Chatelier
91120 Palaiseau 91120
FRANCE

Marvin Pförtner
Universität Tübingen
Fachbereich Informatik
Maria-von-Linden-Strasse 6
72076 Tübingen
GERMANY

Prof. Dr. Giorgia Ramponi
University of Zurich
Binzmühlestrasse 14
8050 Zürich
SWITZERLAND

Prof. Dr. Lorenzo Rosasco
University of Genova and Massachusetts
Institute of Technology and Istituto
Italiano di Tecnologia
via dodecaneso 35
16146 Genova
ITALY

Prof. Dr. Bernhard Schölkopf
Max Planck Institute for Intelligent
Systems
Max-Planck-Ring 4
72076 Tübingen
GERMANY

Dr. Claire Vernade
Universität Tübingen
Fachbereich Informatik
Maria-von-Linden-Straße 6
72076 Tübingen
GERMANY

Dr. Ulrike von Luxburg
Fakultät für Informatik
Universität Tübingen
Maria-von-Linden-Strasse 6
72076 Tübingen
GERMANY

Vaclav Voracek
Universität Tübingen
72074 Tübingen
GERMANY

Prof. Dr. Daniel Walter
Institut für Mathematik
Humboldt-Universität Berlin
10099 Berlin
GERMANY

Prof. Dr. Andrea Walther
Institut für Mathematik
Humboldt-Universität Berlin
Unter den Linden 6
10099 Berlin
GERMANY

Tim Weiland
Tübingen AI Center
Universität Tübingen
Maria-von-Linden-Straße 6
72076 Tübingen
GERMANY

Dr. Jonathan Wenger
Zuckerman Institute
Columbia University
3227 Broadway
New York, NY 10027
UNITED STATES

Prof. Dr. Fanny Yang
Department of Computer Science
ETH Zürich (CAB G 68)
Universitätsstrasse 6
8092 Zürich
SWITZERLAND