2026 PKU Workshop on Mathematical Foundations of Modern Machine Learning

北京大学数学科学学院
School of Mathematical Sciences,Peking University

Jul 3-5, 2026

BEIJING, CHINA

https://ml-math-2026.github.io/

Contents

Organizing Committee

(alphabetic order)

Fanghui Liu, Shanghai Jiao Tong University

Lei Wu, Peking University

Kun Yuan, Peking University

Sponsoring Institutions

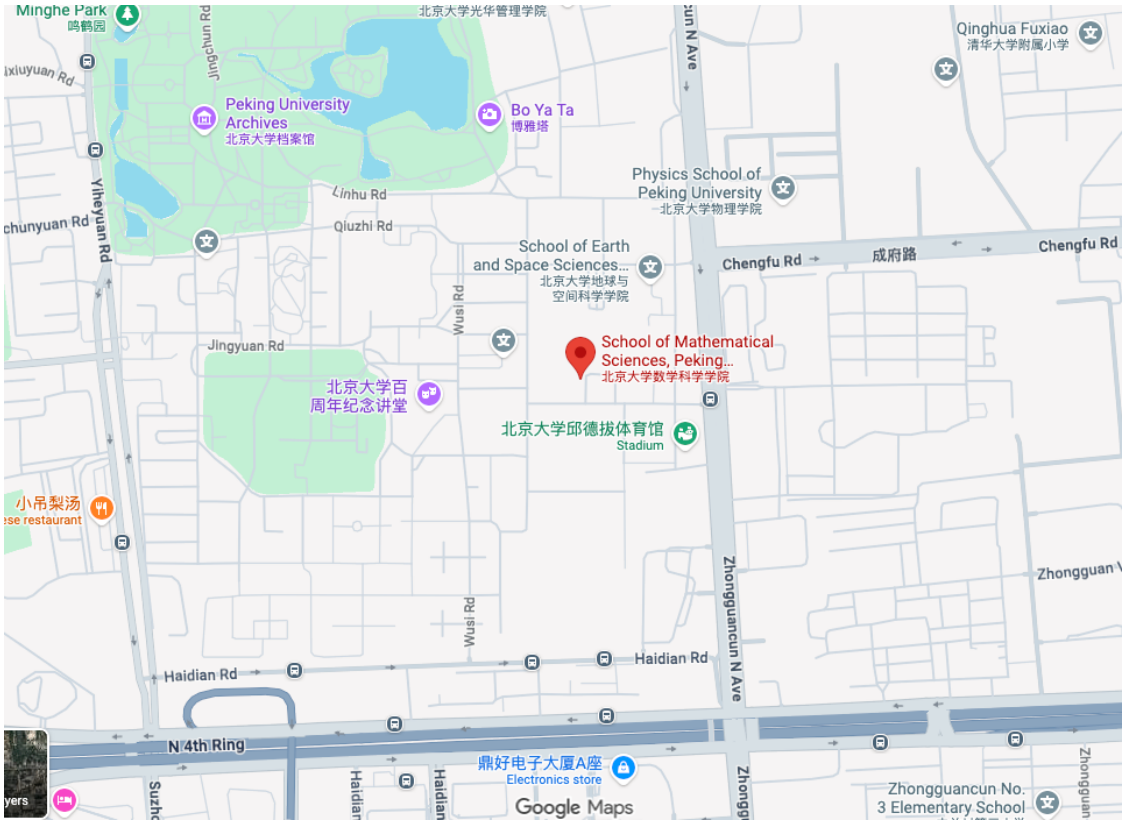
School of Mathematical Sciences, Peking University

Center for Machine Learning Research, Peking University

Institute of Natural Sciences, Shanghai Jiao Tong University

Workshop Location

Ding Shisun 111 Conference Room, Zhihua Building, Peking University



Contact Information

If you need any help, please feel free to contact:

Fanghui Liu, Shanghai Jiao Tong University: fanghui.liu@sjtu.edu.cn

Lei Wu, Peking University: leiwu@math.pku.edu.cn

Kun Yuan, Peking University: kunyuan@pku.edu.cn

Workshop Schedule

Time	Workshop Information	Host
July 3, 2026		
09:00–09:10	Opening Remarks	
09:10–10:00	**Talk: Mikhail Belkin** Feature Learning and the Linear Representation Hypothesis for Steering and Monitoring LLMs	Lei Wu
10:00–10:30	Group Photo & Tea Break	
10:30–11:20	**Talk: Lenka Zdeborová** The Rules-and-Facts Model for Simultaneous Generalization and Memorization in Neural Networks	Yaoyu Zhang
11:20–12:10	**Talk: Gitta Kutyniok** Towards a Theory of Computability for Modern Artificial Intelligence	
12:10–14:30	Lunch	
14:30–15:20	**Talk: Patrick Rebeschini** Generalisation Error and Effective Dimensions: From Neural Networks to Diffusion Models	Jingzhao Zhang
15:20–16:10	**Talk: Mahdi Soltanolkotabi** Steering Generative Models with Verifiable Feedback	
16:10–16:40	Tea Break	
16:40–17:30	**Talk: Lei Wu** Functional Scaling Laws for LLM Pretraining	Kaifeng Lyu
18:00–20:00	Welcome Dinner	
July 4, 2026		
09:00–09:50	**Talk: Jinchao Xu** Neural Networks and Classical Numerical Methods: a Theoretical Perspective	Fanghui Liu
09:50–10:40	**Talk: Volkan Cevher** Optimistic Dual Averaging Unifies Modern Optimizers	
10:40–11:10	Tea Break	

Time	Workshop Information	Host
11:10–12:00	**Talk: Weijie Su** Surrogate-Model Approaches to Optimizers for LLM Training	Zhiqin Xu
12:00–14:10	Lunch	
14:10–15:00	**Talk: Florent Krzakala** A Theory of Deep Learning as Neural Low-Degree Filtering	Hancheng Min
15:00–15:50	**Talk: Fanghui Liu** From SLT4AI to AI4SLT: Generalization under Scaling and Infrastructure in Lean 4	
15:50–16:20	Tea Break	
16:20–17:10	**Talk: Bruno Loureiro** Scaling Laws and Spectra of Shallow Neural Networks in the Feature-Learning Regime	Zhu Li
17:10–18:00	**Talk: Kun Yuan** A Memory Efficient Randomized Subspace Optimization Method for Training Large Language Models	
July 5, 2026		
09:00–09:50	**Talk: Murat A. Erdogdu** Learning Neural Networks in High-Dimensions: Scaling Laws from Regression to Classification	Cong Fang
09:50–10:40	**Talk: Taiji Suzuki** Robust Generalization Ability and Feature Learning Dynamics for Deep Foundation Models: Mean Field Langevin and Associative Recall	
10:40–11:10	Tea Break	
11:10–12:00	**Talk: Denny Wu** Sharp Capacity Scaling of Spectral Optimizers in Learning Associative Memory	Kun Yuan
12:00–12:10	Closing Remarks	

Talks and Abstracts

09:10–10:00, Jul. 3, Friday

Speaker：Mikhail Belkin

Title：Feature Learning and the Linear Representation Hypothesis for Steering and Monitoring LLMs

Abstract： A trained Large Language Model (LLM) contains much of human knowledge. Yet, it is difficult to gauge the extent or accuracy of that knowledge, as LLMs do not always "know what they know" and may even be unintentionally or actively misleading. In this talk I will discuss feature learning, introducing Recursive Feature Machines, a powerful method originally designed for extracting relevant features from tabular data. I will show how this technique enables us to detect and guide LLM behaviors toward almost any desired concept by adding a multiple of fixed vectors in LLM activation spaces. I will give several examples including probing for whether an LLM exhibits motivated reasoning.

Short Bio: Mikhail Belkin is HDSI Endowed Chair Professor in AI at Halicioglu Data Science Institute and Computer Science and Engineering Department at UCSD. From 2023 to 2025 he was an Amazon Scholar. Before moving to UCSD he was a Professor at the Department of Computer Science and Engineering and the Department of Statistics at the Ohio State University. He received his Ph.D. from the Department of Mathematics at the University of Chicago (advised by Partha Niyogi). His research interests are broadly in theory and applications of Artificial Intelligence, deep learning and data analysis.
Some of his well-known work includes widely used Laplacian Eigenmaps, Graph Regularization and Manifold Regularization algorithms, which brought ideas from classical differential geometry and spectral graph theory to data science. His more recent work has been concerned with understanding remarkable mathematical and statistical phenomena observed in deep learning. The empirical evidence necessitated revisiting some of the classical concepts in statistics and optimization, including the basic notion of over-fitting. One of his key findings has been the "double descent" risk curve that extends the textbook U-shaped bias-variance trade-off curve beyond the point of interpolation. His recent work focusses on understanding feature learning and over-parameterization in deep learning.
Mikhail Belkin is an ACM Fellow and a recipient of an NSF CAREER Award and a number of best paper and other awards. He has served on the editorial boards of IEEE Transactions on Pattern Analysis and Machine Intelligence and the Journal of Machine Learning Research. He has served as editor-in-chief of SIAM Journal on Mathematics of Data Science (SIMODS).

10:30–11:20, Jul. 3, Friday

Speaker：Lenka Zdeborová

Title：The Rules-and-Facts Model for Simultaneous Generalization and Memorization in Neural Networks

Abstract： Modern neural networks can both learn general patterns from data and memorize specific facts or exceptions. While both abilities are essential in practice, existing theory largely

treats them separately. In this talk, I will introduce the Rules-and-Facts (RAF) model, a simple framework where part of the data follows an underlying rule, while the rest consists of isolated facts that must be memorized. This setting allows us to study when a learner can simultaneously generalize and recall exceptions. Our results show that the key question is not whether a model has enough capacity, but how this capacity is organized and used. In particular, we identify how regularization and the geometry of the learned representation (e.g., kernel or feature map) determine whether a model can allocate resources to memorize facts without interfering with rule learning. The RAF model thus provides a precise lens on how modern neural networks balance abstraction and memory, and how architectural and algorithmic choices control this trade-off.

Short Bio: Lenka Zdeborová is a Professor of Physics and Computer Science at École Polytechnique Fédérale de Lausanne, where she leads the Statistical Physics of Computation Laboratory. She received a PhD in physics from the University of Paris-Sud and Charles University in Prague in 2008. She spent two years in the Los Alamos National Laboratory as the Director's Postdoctoral Fellow. Between 2010 and 2020, she was a researcher at CNRS, working in the Institute of Theoretical Physics in CEA Saclay, France. In 2014, she was awarded the CNRS bronze medal, in 2016 Philippe Meyer prize in theoretical physics and an ERC Starting Grant, in 2018 the Irène Joliot-Curie prize, in 2021 the Gibbs lectureship of AMS and the Neuron Fund award, in 2025 she received an ERC Advanced Grant. Lenka's expertise is in the application of concepts from statistical physics, such as advanced mean field methods, the replica method, and related message-passing algorithms, to problems in machine learning, signal processing, inference, and optimization. Currently, she focuses on statistical physics of learning, developing solvable models and theoretical principles that explain how modern AI systems generalize, memorize, and scale. She enjoys erasing the boundaries between theoretical physics, mathematics and computer science.

11:20–12:10, Jul. 3, Friday

Speaker：Gitta Kutyniok

Title：Towards a Theory of Computability for Modern Artificial Intelligence

Abstract： Modern AI has achieved remarkable empirical success, yet its fundamental computational capabilities and limitations remain poorly understood. This is particularly true for limitations imposed by the underlying hardware, which have so far received little systematic mathematical study. In this talk, we argue that computability theory offers a promising framework for addressing such questions. We will discuss recent results on computability across digital, analog, and quantum computing paradigms, and outline implications for learning systems, alternative hardware architectures such as neuromorphic computing, and the co-evolution of software and hardware toward a mathematical theory of next-generation AI.

Short Bio: Gitta Kutyniok is Chair for Mathematical Foundations of Artificial Intelligence at LMU Munich, with additional affiliations at the DLR German Aerospace Center and the University of Tromsø. Before joining LMU, she held professorships at the University of Osnabrück and the Technical University of Berlin, where she was an Einstein Chair. She also held visiting positions at various international institutions, including Princeton, Stanford, and Yale University. In 2023, she co-founded EcoLogic Computing GmbH.

Her current research focuses on the mathematical foundations of artificial intelligence, in particular reliable and sustainable AI and automated scientific discovery, with applications in medicine, robotics, and neuroscience.
She has received numerous honors, including a Heisenberg Fellowship, the von Kaven Prize of the DFG, SIAM Fellowship, IEEE Fellowship, and ELLIS Fellowship. She has delivered invited and plenary lectures at major international conferences, including ICM 2022 and ICIAM 2023. She is a member of the Berlin-Brandenburg Academy of Sciences and Humanities and the European Academy of Sciences, Vice President for Science & Research Relations of Venture AI Germany, and LMU Director of the Konrad Zuse School of Excellence in Reliable AI.

14:30–15:20, Jul. 3, Friday

Speaker：Patrick Rebeschini

Title：Generalisation Error and Effective Dimensions: From Neural Networks to Diffusion Models

Abstract： We study generalisation in neural networks and diffusion models using on-average algorithmic stability. For local minimisers of empirical risk, we show that generalisation error can be bounded in terms of data-dependent effective dimensions arising from the geometry of the gradient covariance, capturing low-dimensional structure in the data. For diffusion models, we introduce a notion of score stability and derive generalisation bounds that reflect the role of training and sampling procedures. Across these settings, we study three forms of regularisation: explicit (Tikhonov regularisation), implicit (stochastic gradient descent), and sampling-based (early stopping and discretisation). We show how these control generalisation through different notions of effective dimension.
(Based on joint work with Ayub Kharel, Yasin Abbasi, and Ilja Kuzborskij; and with Tyler Farghly, George Deligiannidis, and Arnaud Doucet)

Short Bio: Patrick Rebeschini serves as a Professor of Statistics and Machine Learning in the Department of Statistics at the University of Oxford. He earned his Ph.D. in Operations Research and Financial Engineering from Princeton University. Before joining Oxford in 2017, he was a postdoctoral researcher in the Departments of Electrical Engineering and Computer Science at Yale University. His research focuses on uncovering and leveraging fundamental principles in high-dimensional probability, statistics, and optimization to develop computationally efficient and statistically optimal algorithms for machine learning and artificial intelligence.

15:20–16:10, Jul. 3, Friday

Speaker：Mahdi Soltanolkotabi

Title：Steering Generative Models with Verifiable Feedback

Abstract： Reinforcement learning with verifiable rewards has become a central tool for improving the reasoning abilities of large language models. Yet many standard RLVR algorithms, including GRPO and RLOO, implicitly emphasize prompts with intermediate

success probability while assigning weak or zero training signal to prompts that are currently very hard. I will first discuss our work on asymmetric prompt weighting, which upweights low-success and even zero-success prompts. This simple modification substantially improves from-scratch RL for reasoning, and our theory characterizes when such asymmetric weighting accelerates progress by reducing the effective time needed to raise success probability.
I will then show how a related principle appears in a seemingly different setting: text-to-image diffusion models. These models often fail to satisfy explicit numerical constraints, such as generating exactly eight objects, because early sampling decisions determine global structure and become difficult to revise later. Our ATHENA framework addresses this by using intermediate count estimates as a lightweight verifiable signal and adaptively steering the denoising trajectory at test time, without retraining or architectural changes. Across both settings, the goal is to use just enough feedback, at just the right time, to keep the model from getting stuck in the wrong trajectory.

Short Bio: Mahdi Soltanolkotabi is the director of the center on AI Foundations for the Sciences (AIF4S) at the University of Southern California. He is also a professor in the Departments of Electrical and Computer Engineering, Computer Science, and Industrial and Systems engineering and part time research scientist at google. Prior to joining USC, he completed his PhD in electrical engineering at Stanford in 2014. He was a postdoctoral researcher in the EECS department at UC Berkeley during the 2014-2015 academic year.
Mahdi is the recipient of the Information Theory Society Best Paper Award, Packard Fellowship in Science and Engineering, an NIH Director's new innovator award, a Sloan Research Fellowship, a Frontiers of Science Award, an NSF Career award, an Airforce Office of Research Young Investigator award (AFOSR-YIP), the Viterbi school of engineering junior faculty research award, and faculty awards from Google and Amazon. His research focuses on developing the mathematical foundations of modern data science via characterizing the behavior and pitfalls of contemporary nonconvex learning and optimization algorithms with applications in AI, deep learning, large scale distributed training, federated learning, computational imaging, and AI for scientific and medical applications. Most recently his applied research is focused on developing and deploying reliable and trustworthy AI in healthcare.

16:40–17:30, Jul. 3, Friday

Speaker：Lei Wu

Title：Functional Scaling Laws for LLM Pretraining

Abstract： Scaling laws are among the most influential empirical regularities in LLM pretraining: final performance improves predictably as data, model size, and compute scale. However, classical scaling laws primarily describe the endpoint of training, leaving open the dynamical mechanism that produces these laws and how scaling behavior is shaped by hyperparameters.
In this talk, I will present Functional Scaling Laws (FSL), a framework that extends scaling laws from final-loss prediction to full loss-trajectory prediction. Starting from power-law kernel regression as a toy model, we derive stochastic training dynamics and reveal an intrinsic-time structure that unifies training curves across scales and protocols.
This functional perspective provides a dynamical theory of scaling laws. It explains how signal learning and noise forgetting jointly determine the loss dynamics, and shows how learning-rate

and batch-size schedules shape scaling behavior through this balance. Experiments on LLM pretraining support FSL as a framework for interpreting and designing large-scale pretraining.

Short Bio: Lei Wu is an Assistant Professor at the School of Mathematical Sciences and the Center for International Machine Learning Research at Peking University. His research focuses on the scientific foundations of deep learning. He received his B.S. in Mathematics from Nankai University in 2012, and his Ph.D. in Computational Mathematics from Peking University in 2018. From November 2018 to October 2021, he conducted postdoctoral research at Princeton University and the University of Pennsylvania. His work has been published in top conferences and journals such as NeurIPS, ICML, COLT, ICLR, Annals of Statistics, Journal of Machine Learning Research, IEEE Transactions on Information Theory.

09:00–09:50, Jul. 4, Saturday

Speaker：Jinchao Xu

Title：Neural Networks and Classical Numerical Methods: a Theoretical Perspective

Abstract： This talk compares neural network-based methods with classical numerical methods from a theoretical perspective. Through several representative examples, we examine both the potential and the limitations of deep neural networks in scientific computing and, more broadly, in machine learning.
We begin by comparing ReLU deep neural networks with polynomials and piecewise polynomial spaces, focusing on their structures and expressive power. We then revisit the curse of dimensionality and discuss whether deep neural networks truly offer advantages over traditional numerical methods for high-dimensional problems.
Next, we consider the use of deep neural networks for solving partial differential equations, with particular emphasis on the challenge of achieving high accuracy.
Finally, we examine multigrid methods and explore whether their underlying principles can help us better understand, design, and train deep neural network models with possible implications for broader AI applications.

Short Bio: Jinchao Xu is Professor of Applied Mathematics and Computational Sciences at KAUST, and Director of the KAUST Lab for Scientific Computing and Machine Learning. His primary research interests include the design, analysis, and application of numerical methods for scientific computing—particularly finite element and multigrid methods—as well as machine learning, including deep neural networks and large language models.
His representative contributions include pioneering theory and algorithms in multigrid and domain decomposition methods; the development of the FASP software package; the formulation of the subspace correction framework; and foundational work on the mathematics of deep learning. He has also contributed to machine learning through MgNet, which bridges ideas from numerical analysis and convolutional neural networks, and AceGPT, an LLM developed specifically for Arabic. Several influential theories and algorithms are named after him (X), including the BPX preconditioner, HX preconditioner, XZ identity, and the MWX element.
He was an invited speaker at the International Congress on Industrial and Applied Mathematics (ICIAM) in 2007 and at the International Congress of Mathematicians (ICM) in 2010. He is a Fellow of the Society for Industrial and Applied Mathematics (SIAM), the American

Mathematical Society (AMS), the American Association for the Advancement of Science (AAAS), the European Academy of Sciences (EURASC), and Academia Europaea.

09:50–10:40, Jul. 4, Saturday

Speaker：Volkan Cevher

Title：Optimistic Dual Averaging Unifies Modern Optimizers

Abstract： At the heart of deep learning's transformative impact lies the concept of scale--encompassing both data and computational resources, as well as their interaction with neural network architectures. Scale, however, presents critical challenges, such as increased instability during training and prohibitively expensive model-specific tuning. Given the substantial resources required to train such models, formulating high-confidence scaling hypotheses backed by rigorous theoretical research has become paramount.
To bridge theory and practice, the talk introduces SODA, a generalization of Optimistic Dual Averaging, which provides a common perspective on state-of-the-art optimizers like Muon, Lion, AdEMAMix and NAdam, showing that they can all be viewed as optimistic instances of this framework. Based on this framing, we propose a practical SODA wrapper for any base optimizer that eliminates weight decay tuning through a theoretically-grounded 1/k decay schedule. Empirical results across various scales and training horizons show that SODA consistently improves performance without any additional hyperparameter tuning.

Short Bio: Volkan Cevher received the B.Sc. (valedictorian) in electrical engineering from Bilkent University in Ankara, Turkey, in 1999 and the Ph.D. in electrical and computer engineering from the Georgia Institute of Technology in Atlanta, GA in 2005. He was a Research Scientist with the University of Maryland, College Park, from 2006-2007 and also with Rice University in Houston, TX, from 2008-2009. He was also a Faculty Fellow in the Electrical and Computer Engineering Department at Rice University from 2010-2020. Currently, he is a Full Professor at the Swiss Federal Institute of Technology Lausanne and an Amazon Scholar. His research interests include machine learning, optimization theory and methods, and automated control. Dr. Cevher is an IEEE Fellow ('24), an ELLIS fellow, and was the recipient of the ICML AdvML Best Paper Award in 2023, Google Faculty Research award in 2018, the IEEE Signal Processing Society Best Paper Award in 2016, a Best Paper Award at CAMSAP in 2015, a Best Paper Award at SPARS in 2009, and an ERC CG in 2016 as well as an ERC StG in 2011.

11:10–12:00, Jul. 4, Saturday

Speaker：Weijie Su

Title：Surrogate-Model Approaches to Optimizers for LLM Training

Abstract： The recent empirical success of the Muon optimizer in training large language models has outpaced the theoretical understanding of its matrix-gradient orthogonalization design. To bridge this gap, this talk introduces surrogate-model approaches that analyze and systematically improve deep learning optimization over a single iteration. We first present the isotropic curvature model, a convex program assuming curvature isotropy across perturbation

directions, which reveals that optimal update matrices achieve a more homogeneous spectrum. This approach demonstrates that while Muon's gradient orthogonalization is directionally correct, it is only strictly optimal under specific curvature phase transitions. Building upon this theoretical foundation, we introduce a second quadratic surrogate model that approximates the loss using the gradient, an output-space curvature matrix, and the input data matrix. By minimizing this surrogate under an isotropic weight assumption, we derive Newton-Muon. This finding implies that standard Muon is an implicit Newton-type method that neglects the right preconditioning induced by the input second moment. Empirically, Newton-Muon accelerates GPT-2 pretraining, reaching target validation loss in 6% fewer iteration steps and reducing wall-clock training time by roughly 4%, illustrating the efficacy of principled surrogate models in designing LLM optimizers. This is based on arXiv:2511.00674 and 2604.01472.

Short Bio: Weijie Su is a professor in the Wharton Statistics and Data Science Department and, by courtesy, in the Departments of Mathematics and Computer and Information Science at the University of Pennsylvania. He serves as a co-director of the Penn Research in Machine Learning (PRiML) Center. Prior to joining Penn, he earned his PhD from Stanford University in 2016 and his bachelor's degree from Peking University in 2011. His research interests include optimization, statistical foundations of generative AI, privacy-preserving data analysis, and high-dimensional statistics. He serves as an associate editor for Operations Research, Journal of Machine Learning Research, Journal of the American Statistical Association, Annals of Applied Statistics, Harvard Data Science Review, and Foundations and Trends in Statistics. He is currently on the ICML 2026 Organizing Committee as the Scientific Integrity Chair and will serve as the Program Chair for ICML 2027.

14:10–15:00, Jul. 4, Saturday

Speaker：Florent Krzakala

Title：A Theory of Deep Learning as Neural Low-Degree Filtering

Abstract： Understanding how deep neural networks learn useful representations from data remains a central challenge in machine learning theory. In this talk, I will present Neural Low-Degree Filtering (Neural LoFi), a stylized limit of gradient-based training in which hierarchical feature learning becomes an explicit spectral procedure. Neural LoFi provides a tractable framework for studying representation learning beyond the lazy regime and offers a concrete mechanism through which depth progressively constructs increasingly complex features from simpler ones. I will discuss the theoretical foundations of the framework, its connection to low-degree methods and kernel learning, and supporting experiments on fully connected and convolutional architectures.
(Based on https://arxiv.org/abs/2605.13612)

Short Bio: Florent Krzakala is Professor of Physics and Electrical Engineering at EPFL, where he leads the IdePHICS laboratory. His research lies at the intersection of statistical physics, machine learning, optimization, and high-dimensional statistics. He is particularly interested in developing theoretical foundations for modern machine learning and understanding the mechanisms underlying representation learning in deep neural networks.

15:00–15:50, Jul. 4, Saturday

Speaker：Fanghui Liu

Title：From SLT4AI to AI4SLT: Generalization under Scaling and Infrastructure in Lean 4

Abstract： This talk connects two directions in modern statistical learning theory. First, I will briefly discuss SLT4AI: how generalization curves behave under suitable model capacities, e.g., norm-based capacities. Our results present generalization under scaling via deterministic equivalence, which challenges classical views of generalization, including the U-shaped bias–variance curve, double descent, and empirical scaling laws.
Second, I will talk about AI4SLT by presenting our Lean 4 formalization of SLT from empirical process theory with more than 30,000 lines code under a human-AI collaborative workflow (with some design principles), including Gaussian Lipschitz concentration, Dudley's entropy integral theorem for sub-Gaussian processes, and application to localized least squares. It highlights how formal AI4SLT infrastructure can make modern learning theory reliable, reusable, and machine-checkable.

Short Bio: Fanghui Liu is currently a (tenure-track) associate professor at Institute of Natural Sciences, School of Mathematical Sciences, Shanghai Jiao Tong University. He previously was an assistant professor at University of Warwick, UK, and a TUM Global Visiting professor. His research interests include mathematical foundations of modern machine learning, mathematical reasoning of LLMs, and AI4Math. He was a recipient of AAAI'24 New Faculty Award, Rising Star in AI (KAUST 2023), co-founded the fine-tuning workshop at NeurIPS'24, and served as an area chair of NeurIPS, ICLR and AISTATS, etc. Besides, he has delivered three tutorials at ISIT'24, CVPR'23, and ICASSP'23, respectively. Prior to his faculty position, he worked as a postdoc researcher at EPFL (2021-2023) and KU Leuven (2019-2023), respectively. He received his PhD degree from Shanghai Jiao Tong University in 2019 with several Excellent Doctoral Dissertation Awards.

16:20–17:10, Jul. 4, Saturday

Speaker：Bruno Loureiro

Title：Scaling Laws and Spectra of Shallow Neural Networks in the Feature-Learning Regime

Abstract： Neural scaling laws underlie many of the recent advances in deep learning, yet their theoretical understanding remains largely confined to linear models. In this work, we present a systematic analysis of scaling laws for shallow neural networks in the feature learning regime. Leveraging connections with matrix compressed sensing and LASSO, we derive a detailed phase diagram for the scaling exponents of the excess risk as a function of sample complexity and weight decay. This analysis uncovers crossovers between distinct scaling regimes and plateau behaviours, mirroring phenomena widely reported in the empirical neural scaling literature. Furthermore, we establish a precise link between these regimes and the spectral properties of the trained network weights, which we characterize in detail. Consequently, we provide a theoretical validation of recent empirical observations connecting the emergence of

power-law tails in the weight spectrum with network generalization performance, yielding an interpretation from first principles.

Short Bio: Bruno Loureiro is a CNRS research scientist at the Computer Science department on École Normale Supérieure in Paris, working on the crossroads of machine learning and statistical mechanics. He holds a PhD degree in Physics from the University of Cambridge, and before moving to ENS he held postdoctoral positions at the École Polytechnique Fédérale de Lausanne (EPFL) and the Institut de Physique Théorique (IPhT) in Paris. He is interested in Bayesian inference, theoretical machine learning and high-dimensional statistics more broadly. His research aims at understanding how data structure, optimisation algorithms and architecture design come together in successful learning.

17:10–18:00, Jul. 4, Saturday

Speaker：Kun Yuan

Title：A Memory Efficient Randomized Subspace Optimization Method for Training Large Language Models

Abstract： The memory challenges associated with training Large Language Models (LLMs) have become a critical concern, particularly when using the Adam optimizer. To address this issue, numerous memory-efficient techniques have been proposed, with GaLore standing out as a notable example designed to reduce the memory footprint of optimizer states. However, these approaches do not alleviate the memory burden imposed by activations, rendering them unsuitable for scenarios involving long context sequences or large mini-batches. Moreover, their convergence properties are still not well-understood in the literature. In this work, we introduce a Randomized Subspace Optimization framework for pre-training and fine-tuning LLMs. Our approach decomposes the high-dimensional training problem into a series of lower-dimensional subproblems. At each iteration, a random subspace is selected, and the parameters within that subspace are optimized. This structured reduction in dimensionality allows our method to simultaneously reduce memory usage for both activations and optimizer states. We establish comprehensive convergence guarantees and derive rates for various scenarios, accommodating different optimization strategies to solve the subproblems. Extensive experiments validate the superior memory and communication efficiency of our method, achieving performance comparable to GaLore and Adam.

Short Bio: Dr. Kun Yuan is an Assistant Professor at Center for Machine Learning Research (CMLR) in Peking University. He completed his Ph.D. degree at UCLA in 2019, and was a staff algorithm engineer in Alibaba (US) Group between 2019 and 2022. His research focuses on optimization and efficient pre-training for LLM. He was the recipient of the 2017 IEEE Signal Processing Society Young Author Best Paper Award, the 2017 ICCM Distinguished Paper Award, and the 2025 IEEE CloudCom Best Paper Runner-Up Award. Some of his work has been integrated to Alibaba MindOpt Solver and NVIDIA DeepStream library.

09:00–09:50, Jul. 5, Sunday

Speaker：Murat A. Erdogdu

Title：Learning Neural Networks in High-Dimensions: Scaling Laws from Regression to Classification

Abstract： We study the optimization and sample complexity of gradient based training of a student neural network in the high-dimensional regime, in both regression and classification settings. We provide a sharp analysis of the SGD dynamics, and derive scaling laws for the risk that highlight the power-law dependences on the optimization time, sample size, and model width.

Short Bio: Murat A. Erdogdu is currently an Associate Professor at the University of Toronto in departments of Computer Science and Statistical Sciences. He is also a faculty member of the Vector Institute, and a CIFAR Chair in AI. Before, he was a postdoctoral researcher at Microsoft Research - New England lab. He obtained his Ph.D. from the Department of Statistics at Stanford University. His research interests include machine learning theory, statistics, applied probability, and connections among these fields.

09:50–10:40, Jul. 5, Sunday

Speaker：Taiji Suzuki

Title：Robust Generalization Ability and Feature Learning Dynamics for Deep Foundation Models: Mean Field Langevin and Associative Recall

Abstract： Feature learning is a fundamental advantage of deep neural network models. While acquiring irrelevant features can severely degrade performance, learning robust representations can enable strong out-of-distribution (OOD) generalization. In this talk, I will explore this phenomenon in two different settings.
In the first half, I consider training a two-layer neural network via mean-field Langevin dynamics. Its long-time limit determines the feature-side geometry learned by the hidden layer. While negative entropy regularization works in the opposite direction to sparsity, our results for Gaussian multi-index problems reveal that the stationary distribution concentrates near the hidden indices, develops a multi-spike structure, and thereby yields parameter recovery with high probability. This results in robust generalization even for out-of-distribution test data. In the second half, I will focus on an associative recall problem, demonstrating that training data diversity allows Transformers to acquire a useful feature called the induction head, which generalizes to OOD test data. It is shown that diversity avoids learning a positional shortcut, and our analysis reveals that there is a clear threshold of diversity required to acquire the induction head. Furthermore, if time permits, I will show that Transformers can handle infinite-length contexts in nonlinear associative recall problems to achieve a minimax optimal rate.

Short Bio: Taiji Suzuki is a Professor in the Department of Mathematical Informatics at The University of Tokyo and a researcher with RIKEN's Center for Advanced Intelligence Project. His work sits at the intersection of machine learning and statistics, with a focus on statistical learning theory, deep learning theory, kernel methods, nonparametric convergence analysis, and optimization; particularly stochastic optimization and optimization for deep learning. Previously, he worked as an Assistant Professor in the Department of Mathematical Informatics, the University of Tokyo, and then he was an Associate Professor in the Department of Mathematical and Computing Science, Tokyo Institute of Technology. He has been a frequent instructor in international summer schools on deep learning theory and

optimization, and served as area chairs of premier conferences such as NeurIPS, ICML, ICLR and AISTATS, a program chair of ACML2019, and an action editor of the Annals of Statistics. He received the Outstanding Paper Award at ICLR in 2021, the JSPS prize, the MEXT Young Scientists' Prize, and Outstanding Achievement Award in 2017 from the Japan Statistical Society.

11:10–12:00, Jul. 5, Sunday

Speaker：Denny Wu

Title：Sharp Capacity Scaling of Spectral Optimizers in Learning Associative Memory

Abstract： Spectral optimizers such as Muon have recently shown strong empirical performance in large-scale language model training. We study their advantage through the linear associative memory problem, a tractable model for factual recall in transformers. In particular, we go beyond orthogonal embeddings and consider Gaussian inputs and outputs, which allows the number of stored associations to exceed the embedding dimension via superposition. Our main result sharply characterizes the recovery rates of one step of Muon, SGD, and Newton's method on the logistic loss under a power law frequency distribution. We show that the storage capacity of Muon significantly exceeds that of SGD, and even matches Newton's method while only using first-order information. Moreover, Muon saturates at a larger critical batch size. We also analyze the multi-step dynamics under a thresholded gradient approximation and show that Muon achieves a faster initial recovery rate than SGD, while both methods eventually converge to the information-theoretic limit. Experiments on synthetic tasks validate the predicted scaling laws.

Short Bio: Denny Wu is an incoming Associate Professor in the Department of Statistics at the University of Oxford. He is currently a Faculty Fellow jointly affiliated with the Center for Data Science at New York University and the Center for Computational Mathematics at the Flatiron Institute. He received his Ph.D. in Computer Science from the University of Toronto and the Vector Institute, where he was supervised by Jimmy Ba and Murat A. Erdogdu. Prior to that, he completed his undergraduate studies at Carnegie Mellon University advised by Ruslan Salakhutdinov.

北京大学用餐指南

以下用餐地点无需使用北大校园卡，支持移动支付。具体开放时间和服务安排可能根据学校情况有所调整，请以现场通知为准。

用餐地点	位置与说明
勺园中餐厅	位置：勺园区域（近留学生宿舍）
勺园西餐厅	位置：勺园区域（近留学生宿舍）
农园食堂（仅三层）	位置：校园中部偏南，二教和图书馆附近 说明：中餐
家园食堂（仅四层）	位置：校园西南部
艺园食堂（仅二层）	位置：校园西南部 说明：中餐
燕南食堂（仅地下）	位置：校园中部偏北，百周年纪念讲堂附近 说明：自选快餐