Session S1.1 (Chair: Taiji Suzuki)

Ding Xuan Zhou	University of Sydney
Title:	Approximation theory of structured deep neural networks

Abstract:

Deep learning has been widely applied and brought breakthroughs in speech recognition, computer vision, natural language processing, and many other domains. The involved deep neural network architectures and computational issues have been well studied in machine learning. But a theoretical foundation for understanding the modelling, approximation or generalization ability of deep learning models with network architectures is still in progress. An important family of structured deep neural networks is deep convolutional neural networks (CNNs) with convolutional structures. The convolutional architecture gives essential differences between deep CNNs and fully-connected neural networks, and the classical approximation theory for fully-connected networks developed around 30 years ago does not apply. This talk describes approximation and generalization analysis of deep CNNs and related structured deep neural networks.

Tomaso Poggio	Massachusetts Institute of Technology

Title: Some principles in the theory of deep learning

Abstract:

In recent years, artificial intelligence researchers have built impressive systems. An important question is whether there exist theoretical principles underlying those architectures, including the human brain, that perform so well in learning tasks. A theory of deep learning could solve many of todays problems around AI, such as explainability and control. Though we do not have a full theory as yet, there are very good reasons to believe in the existence of some fundamental principles of learning and intelligence. I will list some of them and focus on the feature selection properties of SGD.

Session S1.2 (Chair: Sho Sonoda)

Taiji Suzuki	The University of Tokyo / RIKEN AIP
Title:	Non-convex extensions of mean-field gradient methods: applications to reinforcement learning and in-context learning

Abstract:

In the past few years, great progress has been made in the convergence analysis of mean-field Langevin dynamics (MFLD). It is shown that MFLD shows linear convergence for a convex target functional, and its time and space discretization errors are also clarified. However, such a theory cannot be applied to nonconvex objectives. In this presentation, we give some extensions of those theoretical analyses with various potential applications such as reinforcement learning and in-context learning.

First, we discuss applicability of MFLD to reinforcement learning in an actor-critic framework where the policy and Q-function are characterized by mean field neural networks. We show its rate of convergence and derive its time-space discretization error although the estimation of the Q-function includes error and the TD-learning required in the inner loop has a difficulty of convergence stemming from usage of semi-gradient.

Second, we give a convergence analysis of a distributional min-max problem that has a lot of applications such as reinforcement learning. We propose a dual-averaging type method and show its convergence. Although the dual-averaging procedure prevents us from naively applying the log-Sobolev inequality arguments as in the convex setting, we have error bounds which are in a different form than that in the convex setting.

Finally, we show a convergence guarantee of mean field gradient flow (MFGF) for training Transformers to obtain nonlinear features in the pretraining procedure of in-context learning. We show that the objective is strict-saddle and thus the MFGF is not captured by a critical point almost surely. This particularly shows the practical success of in-context learning by Transformers.

Alessandro Sperduti	University of Padova
Title:	A Framework for Flexible Design of Generative Diffusion Models for Graphs

Abstract:

Generative models for graphs can be classified into two prominent families: one-shot models, which generate a graph in one go, and sequential models, which generate a graph by successive additions of nodes and edges. Ideally, between these two extremes lies a continuous range of models that adopt different levels of sequentiality. I will present a graph generative framework, based on the theory of Denoising Diffusion Probabilistic Models (DDPM), that supports the specification of a sequentiality degree. Performances of models designed by this framework for different sequentiality degrees are then presented, in terms of quality, run time, and memory.

Session S1.3 (Chair: Masaaki Imaizumi)

Marco Cuturi	Apple / CREST-ENSAE
Title:	On Neural Transport Methods
Abstract:	
I will present in this work an overview of recent advances in NN modeling of optimal transport maps, as well as applications to the Brenier polar factorization theorem.	
Ichiro Takeuchi	Nagoya University / RIKEN AIP
Title:	Statistical Test for Deep Learning-Driven Hypotheses
Abstract:	
When utilizing hypotheses generated by deep learning models in high-stakes decision-making tasks, such as medical diagnosis or autonomous driving, it is essential to ensure their reliability. In this talk, I introduce an approach to evaluate the reliability of deep learning-driven hypotheses based on statistical tests. Since deep learning generates hypotheses through complex transformations of intricate data, developing valid statistical tests is considered challenging. However, by employing a new statistical inference framework known as Selective Inference, we can obtain valid p-values with theoretical guarantees, even with finite sample sizes. In this talk, I will showcase examples where p-values for hypotheses generated by various deep learning models, including CNNs, Transformers, and Diffusion Models, have proven effective in medical image analysis tasks.	
Kenji Fukumizu	The Institute of Statistical Mathematics
Title:	Extended Flow Matching for Conditional Generation
Abstract:	
The task of conditional generation is one of the most important applications of generative models, and numerous methods have been developed to date based on the celebrated diffusion models, with the guidance-based classifier-free method taking the lead. However, the theory of the guidance-based method not only requires the user to fine-tune the ``guidance strength'', but its target vector field does not necessarily correspond to the conditional distribution used in the training. In this work, we develop the theory of conditional generation based on Flow Matching, a current strong contender of diffusion methods. Motivated by the interpretation of a probability path as a distribution on path space, we establish a novel theory of flow-based generation of conditional distribution by employing the mathematical framework of generalized continuity equation instead of the continuity equation in flow matching. This theory naturally derives a method that aims to match the matrix field instead of the vector field. Our framework ensures the continuity of the generated conditional distribution through the flow between conditional distributions. We will present our theory through experiments and mathematical results.	

Session S1.4 (Chair: Masashi Sugiyama)

Klaus-Robert Mueller	Technische Universität Berlin
Title:	ML meets Quantum Chemistry
Abstract:	

The introduction of ML in Quantum Chemistry has created a new research direction and has given rise to significant and insightful progress. The talk first introduces the topic and subsequently discusses recent developments. Emphasis is placed at a reflection of mutual cross-fertilization; e.g. physics challenges gave rise to novel ML architectures.

Naoto Yokoya	The University of Tokyo / RIKEN AIP
Title:	Advancing Remote Sensing with Deep Learning
Abstract:	

Deep learning has led to remarkable advances in remote sensing data analysis, improving both the accuracy and feasibility of previously difficult tasks. However, a notable challenge remains: the scarcity of large-scale training data for many remote sensing tasks. In this talk, we will highlight our recent initiatives aimed at mitigating this challenge in the understanding and acquisition of remote sensing imagery. Specifically, we'll showcase deep learning methods tailored for land cover semantic recognition and topographic information extraction from remote sensing data. Additionally, we'll present advances in computational imaging designed to overcome the inherent limitations of spatial, temporal, and spectral resolution in imaging systems.

Le Song	Biomap / MBZUAI
Title:	Foundational AI Models for Biological Systems
Abstract:	

What will be the foundational AI models for biological systems? What data can be used to build them? How to build them exactly? Nowadays, biological data grow rapidly and converge into a few standard modalities, such as DNA, RNA and protein sequences and structures, biomolecular interaction networks, and single-cell RNA sequencing and imaging. It seems timely to ask the intriguing questions as to whether foundational AI models can be established for biological systems which possess certain level of generality and transferability and can serve as the infrastructure to enhance the entire spectrum of downstream prediction tasks from different scales of biological systems.

In this talk, I will share my recent work along this direction and introduce the xTrimo family of large scale pretrained models leveraging a large amount of data from protein sequences, structures, protein-protein interactions, and single-cell transcriptomics. The pretrained models can be used as the foundation to address many predictive tasks arising from protein design and cellular engineering and achieve SOTA performances.

Session S2.1 (Chair: Pierre Baldi)

Shun-ichi Amari	RIKEN
Title:	Wasserstein Statistics and AI

Abstract:

Information geometry and Wasserstein geometry provide different geometrical structures to the manifold of probability distributions. Both are used in the studies of statistical inference and deep learning. We focus here on the characteristics of Fisher-based statistics and Wasserstein-based statistics. Given an empirical distribution from observed data and a statistical model, F-estimator is the minimizer of the KL-divergence and W-estimator is the minimizer of the W distance from the empirical distribution to the model. The F-efficiency and W-efficiency are discussed.

Nihat Ay	Hamburg University of Technology
Title:	On the Fisher-Rao gradient of the Evidence Lower Bound

Abstract:

The talk is based on joint work with Jesse van Oostrum and Adwait Datar. In this work, we study the Fisher-Rao gradient, also referred to as the natural gradient, of the evidence lower bound, the ELBO, which plays a crucial role within the theory of the Variational Autonecoder, the Helmholtz Machine and the Free Energy Principle. The natural gradient of the ELBO is related to the natural gradient of the Kullback-Leibler divergence from a target distribution, the primary objective function of learning. Based on invariance properties of gradients within information geometry, conditions on the underlying model are provided that ensure the equivalence of minimising the primary objective function and the maximisation of the ELBO.

Minh Ha Quang	RIKEN AIP
Title:	Infinite-dimensional Fisher-Rao metric and Wasserstein distances

Abstract:

Information geometry and Optimal Transport have been attracting much research attention in various fields, in particular machine learning and statistics. In this talk, we present results on the generalization of the central concepts of Fisher-Rao metric and Wasserstein distances for finite-dimensional Gaussian measures to the setting of infinite-dimensional Gaussian measures and Gaussian processes. The mathematical formulation involves the interplay of Information Geometry, Optimal Transport, and Operator Theory, along with the theory of Gaussian processes and the methodology of reproducing kernel Hilbert spaces (RKHS).

Session S2.2 (Chair: Shun-ichi Amari)

Pierre Baldi	University of California Irvine
Title:	A Theory of Neuronal Synaptic Balance

Abstract:

When a typical feed-forward neural network is trained by gradient descent, with an L2 regularizer to avoid overly large synaptic weights, a strange phenomenon occurs: at the optimum, each neuron becomes "balanced" in the sense that the L2 norm of its incoming synaptic weights becomes equal to the L2 norm of its outgoing
synaptic weights. We develop a theory that explains this phenomenon and exposes its generality. Balance emerges with a variety of activation functions, a variety of regularizers including all Lp regularizers, and a variety of networks including recurrent networks. A simple local balancing algorithm can be applied to any neuron and at any time, instead of just at the optimum.
Most remarkably, for any starting point, stochastic iterated application of the local balancing algorithm always converges to a unique, globally balanced, state.

Emtiyaz Khan	RIKEN AIP
Title:	The Bayesian Learning Rule

Abstract:

Humans and animals have a natural ability to autonomously learn and quickly adapt to their surroundings. How can we design machines that do the same? In this talk, I will present Bayesian principles to bridge such gaps between humans and machines. I will show that a wide-variety of machine-learning algorithms are instances of a single learning-rule derived from Bayesian principles. I will show our recent result on scaling up variational learning to large deep networks (e.g., GPT-2). Time permitting, I will also briefly discuss the dual perspective yielding new mechanisms for knowledge transfer in learning machines.

Session S2.3 (Chair: Emtiyaz Khan)

Kenji Doya	Okinawa Institute of Science and Technology
Title:	Bayesian inference, reinforcement learning, and the cortico-basal ganglia circuit

Abstract:

Bayesian inference is a standard way of handling uncertainties in sensory perception and reinforcement learning is a common way of acting in unknown environments. While they are used in combination for perception and action in uncertain environments, the similarity of their computations has been formulated as the duality of inference and control, or control as inference.
In this talk, I will review these theoretical frameworks and discuss their implications in understanding the common circuit architectures of the sensory and motor cortices, and possible roles of the basal ganglia in motor and sensory processing.

Ilsang Ohn	Inha University
Title:	On adaptive inference with variational Bayes

Abstract:

In this work, we propose a novel approach for adaptive inference with variational Bayes. The proposed method first computes a variational posterior over each individual model separately and then combines them with certain weights to produce a variational posterior over the entire model. We show that this aggregated variational posterior can be a good approximation to the original posterior over the entire model under mild conditions, and due to this approximation property, it can attain adaptive contraction rates. We illustrate the general results in a number of examples, including deep neural networks and Gaussian processes.

Yongdai Kim	Seoul National University
Title:	Posterior concentration rates of Bayesian deep neural networks

Abstract:

Bayesian methods in training deep neural networks (BNNs) have received much attention and have been effectively utilized in a wide range of AI applications
In this talk, theoretical properties of BNNs are considered. In particular, posterior concentration rates are derived for regression problems where
the true regression function is smooth. Firstly, existing results about the posterior concentration rates or BNN are reviewed and their limitations are discussed. Then, a new result of the posterior concentration rates is given which reduces a gap between theories and applications. If time is allowed, some issues of computations in BNNs are discussed.

Session S2.4 (Chair: Kenji Fukumizu)

Sho Sonoda	RIKEN AIP
Title:	Ridgelet Transform: Harmonic Analysis for Deep Neural Networks
Abstract:	

Ridgelet transform is a pseudo-inverse operator of neural networks. Namely, given a function $f ¥in L^2(¥mathbb{R}^m)$, the ridgelet transform $R[f]$ describes how the network parameters should be distributed for the network to represent f. In this talk, I will explain two systematic schemes to derive the ridgelet transform. As applications, we investigate modern neural networks involving the ones on manifolds G/K and Hilbert spaces H as well as deep networks, and derive their associated ridgelet transforms.

Masaaki Imaizumi	The University of Tokyo / RIKEN AIP
Title:	Statistical Analysis on In-Context Learning
Abstract:	

Deep learning and artificial intelligence technologies have made great progress, and the usage of foundation models has attracted strong attension by its general ability. Motivated by this fact, mathematical understanding is required to efficiently control and develop these technologies. In this talk, I will present a statistics-based analysis of a scheme called in-context learning, which is an useful framework of meta-learning to describe foundation models. I argue that in-context learning can efficiently learn the latent structure of the data, using the property of transformers used in the learning scheme can efficiently handle the distribution of observations.

Frank Wang	National Taiwan University & NVIDIA
Title:	Vision Language Models for Novel Object Captioning & Natural Language Explanation
Abstract:	

The convergence of language, vision, and generative models is a captivating and rapidly advancing research domain. In this talk, we will delve into the intricate interplay between these disciplines, showcasing how generative models have sparked a revolution in creative and analytical applications. We will explore the mechanisms behind models' ability to decode images into text and vice versa, shedding light on their potential to reshape human-machine interaction. With the introduction of a number of our recent research works on novel object captioning and natural language explanation, we will also discuss its challenges and emerging opportunities.

Session S3.1 (Chair: Ding Xuan Zhou)	
Sumio Watanabe	Tokyo Institute of Technology
Title:	Singular Learning Theory, Deep Learning, and AI Alignment
Abstract:	
From the viewpoint of statistical learning theory, deep neural networks are nonlinear and nonregular statistical models. Neither the likelihood function nor the posterior distribution of them can be approximated by any normal distribution, resulting that algebro-geometric foundation is necessary to study their generalization performance. In this talk, we explain singular learning theory by which both the marginal likelihood and the generalization error are represented by birational invariants, and introduce several methods how to apply them to AI alignment.	
Guido Montúfar	University of California, Los Angeles
Title:	Mildly Overparameterized ReLU Networks Have a Favorable Loss Landscape
Abstract:	
We study the loss landscape of both shallow and deep, mildly overparameterized ReLU neural networks on a generic finite input dataset for the squared error loss. We show both by count and volume that most activation patterns correspond to parameter regions with no bad local minima. Furthermore, for one-dimensional input data, we show most activation regions realizable by the network contain a high dimensional set of global minima and no bad local minima. We experimentally confirm these results by finding a phase transition from most regions having full rank Jacobian to many regions having deficient rank depending on the amount of overparameterization. The talk is based on work with Kedar Karhadkar, Michael Murray, and Hanna Tseran.	
Atsushi Nitanda	A*STAR
Title:	Convergence Analysis of Mean-field Optimization
Abstract:	
Optimization of mean-field models has recently attracted attention due to its connection to training two-layer neural networks under the mean-field regime. We have established the theory of convex analysis for these models to analyze their optimization dynamics and have derived various optimization methods with convergence guarantees. In this talk, I present recent advances in mean-field optimization methods including mean-field Langevin dynamics.	
Session S3.2 (Chair: Qibin Zhao)	
Akiko Takeda	The University of Tokyo / RIKEN AIP
Title:	Random subspace optimization methods for large-scale optimization problems
Abstract:	
Random projection techniques based on the Johnson-Lindenstrauss lemma are used for randomly aggregating the constraints or variables of optimization problems while approximately preserving their optimal values, which leads to smaller-scale optimization problems. In this talk, we show a few applications of random matrix techniques for constructing random subspace algorithms that iteratively solve smaller-scale subproblems and discuss their convergence speed. This talk is based on joint works with Ryota Nozawa, and Pierre-Louis Poirion.	
Amy Zhang	The University of Texas at Austin
Title:	Self-supervised Reinforcement Learning for Zero-shot Optimality
Abstract:	
Recent work in language and video generation have shown the benefits of autoregressive objectives for generation of sequential data. However, these methods also often require additional fine-tuning, like RLHF, in order to output the desired behavior, especially if the data quality is poor. In this talk, we explore a different paradigm borrowed from work on self-supervised RL, or zero-shot RL. We will present methods that only assume access to unlabeled sequential data for which we can provably extract the optimal value function given access to task information. We first explore the goal-conditioned RL setting, which allows us to extract optimal value functions for goal states, then an expanded setting that can handle any downstream task.	

Session S3.3 (Chair: Amy Zhang)

Han Zhao	University of Illinois at Urbana-Champaign
Title:	Revisiting Scalarization in Multi-Task Learning: A Theoretical Perspective

Abstract:

Linear scalarization, i.e., combining all loss functions by a weighted sum, has been the default choice in the literature of multi-task learning (MTL) since its inception. In recent years, there has been a surge of interest in developing Specialized Multi-Task Optimizers (SMTOs) that treat MTL as a multi-objective optimization problem. However, it remains open whether there is a fundamental advantage of SMTOs over scalarization. In fact, heated debates exist in the community comparing these two types of algorithms, mostly from an empirical perspective. In this talk, I will revisit scalarization from a theoretical perspective. I will be focusing on linear MTL models and studying whether scalarization is capable of fully exploring the Pareto front. Our findings reveal that, in contrast to recent works that claimed empirical advantages of scalarization, scalarization is inherently incapable of full exploration, especially for those Pareto optimal solutions that strike the balanced trade-offs between multiple tasks. More concretely, when the model is under-parametrized, we reveal a multi-surface structure of the feasible region and identify necessary and sufficient conditions for full exploration. This leads to the conclusion that scalarization is in general incapable of tracing out the Pareto front. Our theoretical results provide a more intuitive explanation of why scalarization fails beyond non-convexity.

Arthur Gretton	University College London
Title:	Learning to act in noisy contexts using deep proxy learning

Abstract:

We consider problem of evaluating the expected outcome of an action or policy, using off-policy observations of user actions, where the relevant context is noisy/anonymized. This scenario might arise due to privacy constraints, data bandwidth restrictions, or intrinsic properties of the setting.

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 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.

Masashi Sugiyama	RIKEN AIP / The University of Tokyo
Title:	Importance-Weighting Approach to Distribution Shift Adaptation

Abstract:

Standard machine learning methods suffer from distribution shifts between training and test data. In this talk, I will first give an overview of the classical importance weighting approach to distribution shift adaptation, which consists of an importance estimation step and an importance-weighted training step. Then, I will present a more recent approach that simultaneously estimates the importance weight and trains a predictor. I will also discuss a more practical scenario of continuous distribution shifts, where the data distributions change continuously over time. Finally, I will discuss ongoing challenges such as joint distribution shift and out-of-distribution adaptation.

Session S3.4 (Chair: Han Zhao)

Qibin Zhao	RIKEN AIP
Title:	Efficient and Robust Machine Learning with Tensor Networks

Abstract:

Modern ML methods have achieved the remarkable performance by dramatically increasing the DNN model size and the amount of high quality data samples. However, how to learn information from data efficiently and train a parameter efficient model become important in particular applications. Tensor Networks (TNs) have been increasingly investigated and applied to machine learning and signal processing, due to their advantages in handling large-scale and high-dimensional problems, model compression in DNNs, and efficient computations for learning algorithms. This talk aims to present some recent progresses of TNs technology applied to machine learning from perspectives of basic principle and algorithms, particularly in unsupervised learning, data completion, multi-model learning and various applications in deep learning modeling and adversarial robustness. Finally, we will also present potential research directions and new trends in this area.