2024 Workshop on Mathematical Theories and Algorithms for AI for Science

Date: September 19-21, 2024

Organizers:

Zhichao Peng, The Hong Kong University of Science and Technology

Yang Xiang, The Hong Kong University of Science and Technology

Kun Xu, The Hong Kong University of Science and Technology

Can Yang, The Hong Kong University of Science and Technology

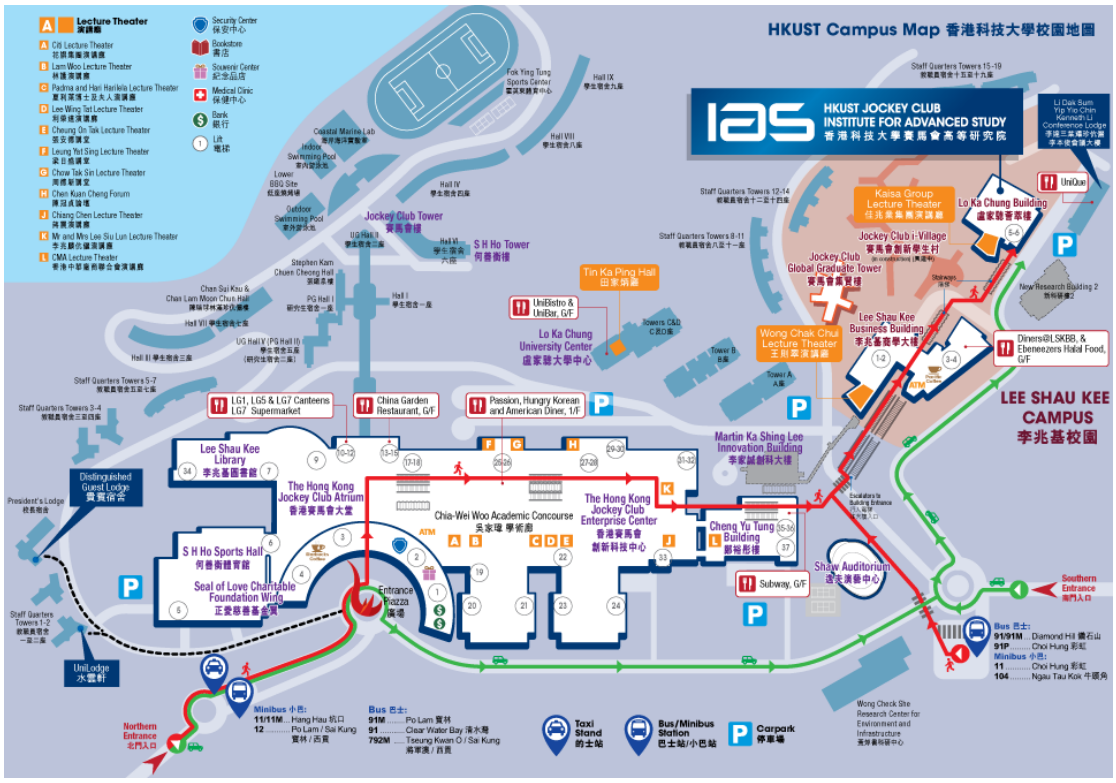
Sponsored by

Department of Mathematics, HKUST and Huawei

Location:

HKUST Jockey Club, Institute for Advanced Study (IAS), Room 1038

Traffic information: https://cso.ust.hk/tran/pt



Day 1, 19/09/2024

Time	Speaker	Talk
09:20-09:30	Opening	
09:30-10:15	Lei Wu	Kernel Regression and Random Feature Approximation: Equivalence and Saturation
10:15-11:00	Fenglei Fan	Hyper-Compression: Neural Network Compression via Hyper-function
11:00-11:20	Coffee break	
11:20-12:05	Yuan Cao	Understanding Over-Parameterized Learning: Phenomena and Explanations
12:05-14:00	Lunch break	
14:00-14:45	Gen Li	Faster Convergence and Acceleration for Diffusion-Based Generative Models
14:45-15:30	Rong Tang	Adaptivity of Diffusion Models to Manifold Structures
15:30-15:50	Coffee break	
15:50-16:35	Tianyu Wang	Three Challenges in Non-Euclidean Learning
16:35-17:20	Shuang Qiu	Traversing Pareto Optimal Policies: Provably Efficient Multi-Objective Reinforcement Learning
17:30-20:00	Banquet	

Day 2, 20/09/2024

Time	Speaker	Talk
09:15-10:00	Ying Jiaxi	Inferring Graphs from Data: Algorithms and Theory
10:00-10:45	Zihao Hu	Optimization on Geodesic Metric Spaces: Some Examples
10:45-11:00	Coffee break	
11:00-11:45	Tao Luo	The Theory of Parameter Condensation in Neural Networks
11:45-12:30	Zhichao Peng	A learning-based projection method for model order reduction of transport problems
12:30-14:00	Lunch break	
14:00-18:00	Free discussion	

Day 3, 21/09/2024

Time	Speaker	Talk
9:00-16:00	Free Discussion	

Dr. Lei Wu (PKU, Thursday 9:30-10:15)

BIO: Dr. Lei Wu is currently an assistant professor in the School of Mathematical Sciences and Center for Machine Learning Research at Peking University. His research centers on theoretical aspects of deep learning. He received his Bachelor's degree in pure mathematics from Nankai University in 2012 and a PhD degree in computational mathematics from Peking University in 2018. From November 2018 to October 2021, he worked as a postdoctoral researcher separately at Princeton University and the University of Pennsylvania.

Title: **Kernel Regression and Random Feature Approximation: Equivalence and Saturation**

Abstract:

In this talk, we provide a comprehensive analysis of the sample complexity of kernel ridge regression (KRR) in the noiseless regime, a scenario crucial to scientific computing, where data are often generated via computer simulations. We establish that up to logarithmic terms, KRR can attain the minimax optimal rate, which can adapt to both the eigenvalue decay of the associated integral operator and the relative smoothness of target functions. Remarkably, this rate can be significantly faster than the Monte Carlo rate due to the adaptivity to the target function’s structure. Our analysis is based on an equivalence between KRR and random feature approximation, which may be of independent interest. Additionally, we provide numerical experiments that support our theoretical findings.

Dr. Fenglei Fan (CUHK, Thursday 10:15-11:00)

Bio: Dr. Fenglei Fan is currently a Research Assistant Professor in the Department of Mathematics at the Chinese University of Hong Kong. His research focuses on mathematical AI. He obtained his Ph.D. from Rensselaer Polytechnic Institute in the United States in 2021. He then conducted a one-year postdoctoral research at Cornell University. His doctoral dissertation won the 2021 Outstanding Doctoral Dissertation Award by the International Neural Network Society (INNS). His representative work was selected as one of few 2024 CVPR Best Paper Award candidates (26 out of 1W+ submissions), and won IEEE TRPMS Best Paper Award. As an RAP, he leads one Huawei Gifted Fund and one key research project from Huawei.

Title: **Hyper-Compression: Neural Network Compression via Hyper-function**

Abstract:

Recently, the escalating demand for memory and computational resources by large models has presented formidable challenges for their deployment in resource-constrained environments. One prevalent way to serve large models is to develop effective model compression approaches that crop the size of large models while maintaining acceptable performance levels. Currently, the landscape of model compression methodologies predominantly revolves around four fundamental algorithms: pruning, quantization, knowledge distillation, and low-rank approximation whose basic frameworks have been established years ago. However, the compression efficacy of these methods is often capped or challenging to scale based on theoretical analysis. In this talk, we introduce a novel and general-purpose approach, referred to as hyper-compression, that redefines model compression as a problem of parameter representation. Specifically, we extend the concept of hypernets into what we term a 'hyperfunction'. Then, the hyperfunction is designed based on ergodic theory (ET). This advanced formulation leads to a performant algorithm that offers several distinct advantages, succinctly summarized as **PNAS**: 1) **P**referable compression ratio; 2) **N**o post-hoc retraining; 3) **A**ffordable inference time; and 4) **S**hort compression time. Lastly, in light of the observed stagnation in hardware Moore's Law, we conjecture "Moore's Law of Model Compression", *i.e.*, the efficiency of model compression could double annually in the near future to meet the needs of large model era. We believe that model compression based on hyperfunction can play an important role in "Moore's Law of Model Compression".

Dr. Yuan Cao (HKU, Thursday, 11:20-12:05)

Bio: Dr. Yuan Cao is an assistant professor in the Department of Statistics and Actuarial Science and Department of Mathematics at the University of Hong Kong. Before joining HKU, he was a postdoctoral scholar at UCLA. He received his B.S. from Fudan University and Ph.D. from Princeton University. Yuan's research interests include deep learning theory, non-convex optimization, and high-dimensional statistics. He has published research papers in top machine learning journals (JMLR, ML, TMLR, etc) and conferences (NeurIPS, ICML, ICLR, COLT, AAAI, IJCAI, etc).

Title: Understanding Over-Parameterized Learning: Phenomena and Explanations

Abstract:
Deep learning has achieved great success in many applications. However, the success of deep learning has not been well understood in theory. A key challenge for the theoretical studies of deep learning is the fact that modern neural networks are often highly over-parameterized. In this talk, I will delve into several recent theoretical investigations on over-parameterized learning. Specifically, I will present theoretical results on topics such as "benign overfitting", the influence of optimization algorithms on learning outcomes, and the intriguing "double/multiple descent" phenomenon. Notably, the proofs of these theoretical results are all highly motivated by phenomena observed through carefully designed experiments.

Dr. Gen Li (CUHK, Thursday 14:00-14:45)

Bio: Gen Li is an assistant professor in the Department of Statistics at the Chinese University of Hong Kong (CUHK). Before joining CUHK, he was a postdoc at University of Pennsylvania, and obtained his Ph.D. and B.S. in the Department of Electronic Engineering at Tsinghua University. His research interest includes diffusion based generative model, reinforcement learning, high-dimensional statistics, machine learning, signal processing, and mathematical optimization.

Title: Faster Convergence and Acceleration for Diffusion-Based Generative Models

Abstract:

Diffusion models, which generate new data instances by learning to reverse a Markov diffusion process from noise, have become a cornerstone in contemporary generative modeling. While their practical power has now been widely recognized, theoretical underpinnings for mainstream samplers remain underdeveloped. Moreover, despite the recent surge of interest in accelerating diffusion-based samplers, convergence theory for these acceleration techniques remains limited. In this talk, I will introduce a new suite of non-asymptotic results aimed at better understanding popular samplers like DDPM and DDIM in discrete time, offering significantly improved convergence guarantees over previous work. Our theory accommodates L2-accurate score estimates, and does not require log-concavity or smoothness on the target distribution. Building on these insights, we propose training-free algorithms that provably accelerate diffusion-based samplers, leveraging ideas from higher-order approximation similar to those used in high-order ODE solvers like DPM-Solver. Our acceleration algorithms achieve state-of-the-art sample quality compared to existing methods.

Dr. Rong Tang (HKUST, Thursday 14:45-15:30)

Bio: Dr. Tang is an assistant professor in the Department of Mathematics at the Hong Kong University of Science and Technology (HKUST) since July, 2023. Prior to that, she received her Ph.D. in Statistics from the University of Illinois at Urbana-Champaign (UIUC), advised by Prof. Yun Yang, and completed her B.S. in Statistics at Zhejiang University in 2018.

Title: Adaptivity of Diffusion Models to Manifold Structures

Abstract:
Empirical studies have demonstrated the effectiveness of (score-based) diffusion models in generating high-dimensional data, such as texts and images, which typically exhibit a low-dimensional manifold nature. These empirical successes raise the theoretical question of whether score-based diffusion models can optimally adapt to low-dimensional manifold structures. While recent work has validated the minimax optimality of diffusion models when the target distribution admits a smooth density with respect to the Lebesgue measure of the ambient data space, these findings do not fully account for the ability of diffusion models in avoiding the curse of dimensionality when estimating high-dimensional distributions. This work considers two common classes of diffusion models: Langevin diffusion and forward-backward diffusion. We show that both models can adapt to the intrinsic manifold structure by showing that the convergence rate of the inducing distribution estimator depends only on the intrinsic dimension of the data. Moreover, our considered estimator does not require knowing or explicitly estimating the manifold. We also demonstrate that the forward-backward diffusion can achieve the minimax optimal rate under the Wasserstein metric when the target distribution possesses a smooth density with respect to the volume measure of the low-dimensional manifold.

Dr. Tianyu Wang (Fudan University, Thursday 15:50-16:35)

Bio: Dr. Tianyu Wang is currently an associate young investigator (tenure-track) at Shanghai Center for Mathematical Sciences, Fudan University. Before joining Fudan, Dr. Wang obtained his Ph.D. from Duke University and his Bachelor's degree from HKUST.

Title: Three Challenges in Non-Euclidean Learning

Abstract:
In this talk, I will discuss three challenges in non-Euclidean learning: the challenge for global learning, the challenge for specifying local coordinate systems, and the challenge for analyzing iterative algorithms. We will explore potential solutions for these challenges and demonstrate their effectiveness through important examples.

Dr. Shuang Qiu (HKUST, Thursday 16:35-17:20)

Bio: Dr. Shuang Qiu is currently a Research Assistant Professor in the Department of Mathematics at the Hong Kong University of Science and Technology. He earned his Ph.D. in Computer Science and Engineering from the University of Michigan in 2021. He was a postdoctoral researcher in the Booth School of Business at the University of Chicago. His research primarily focuses on developing provably efficient sequential decision-making methods that lie at the intersection of reinforcement learning, game theory, stochastic optimization, and statistical learning. His research interests also extend to the practical applications of decision-making methods across various research fields, such as foundation models and embodied AI.

Title: Traversing Pareto Optimal Policies: Provably Efficient Multi-Objective Reinforcement Learning

Abstract:

Multi-objective reinforcement learning (MORL) has gained significant attention in recent years due to its impressive performance in various fields such as robotics, intelligent control, business, healthcare, and the prevailing large language models. MORL focuses on learning all Pareto optimal policies in the presence of multiple different or even conflicting reward functions, a more challenging scenario than single-objective RL. Despite MORL's great empirical success, there remains a lack of satisfactory understanding of MORL from a theoretical perspective. This talk first offers a systematic analysis of several optimization targets for MORL, evaluating their abilities to find all Pareto optimal policies and controllability over learned policies based on preferences for different objectives. And a class of Tchebycheff scalarization methods are further identified as favorable optimization targets for MORL. Considering the challenge of minimizing these targets directly, we then reformulate their minimization problems into novel min-max-max optimization problems. Finally, we present novel upper confidence bound (UCB)-based preference-aware and preference-free algorithms with provable convergence guarantees to efficiently learn the Pareto optimal policies under different preferences, with the preference-free algorithm significantly saving exploration cost.

Dr. Jiaxi Ying (HKUST, Friday 9:15-10:00)

Bio: Dr. Jiaxi Ying is currently a postdoctoral fellow in the Department of Mathematics at the Hong Kong University of Science and Technology, working with Prof. Jian-feng Cai. Prior to this, he completed his Ph.D. in the Department of Electronic and Computer Engineering from the same university in 2022. He is broadly interested in developing algorithms that are computationally and statistically efficient to address problems in machine learning, signal processing, and network science utilizing tools from optimization, statistics, and information theory.

Title: Inferring Graphs from Data: Algorithms and Theory

Abstract:

In an era of big data, inferring meaningful relationships from complex datasets is crucial across various fields. This presentation explores recent advancements in inferring graphs from data using graphical models, focusing on two novel approaches that address computational challenges in high-dimensional settings. First, we introduce an innovative algorithm based on the two-metric projection method for solving a sign-constrained log-determinant program. This second-order algorithm enhances convergence rates while maintaining the same per-iteration computational complexity as the gradient projection method. Second, we present a decomposition strategy that leverages the concept of "bridges" in large-scale sparse graphs, allowing the graph inference problem to be divided into smaller, more manageable sub-problems with explicit solutions for bridge-related entries. This approach significantly boosts computational efficiency and scalability.

Dr. Zihao Hu (HKUST, Friday 10:00-10:45)

Bio: Dr. Zihao Hu is currently a research assistant professor at the Department of Mathematics, Hong Kong University of Science and Technology. Previously, he got his Ph.D. from Georgia Institute of Technology, supervised by Professor Jacob Abernethy. His research interests lie in online decision-making theory and mathematical optimization with a geometric structure. Recently, he has been exploring the application of online learning theory to real-world problems, such as online first-price auctions.

Title: Optimization on Geodesic Metric Spaces: Some Examples

Abstract:
There has been growing interest in designing optimization algorithms with convergence guarantees when the parameter space is not a Euclidean set but a Riemannian manifold. This has attracted considerable attention because it allows for the encoding of constraints, and in many cases, it addresses the non-convexity of both the feasible set and the objective function. This talk is divided into two parts. In the first part, we consider projection-free (online) optimization on Riemannian manifolds. We illustrate how to design algorithms that rely solely on a linear optimization oracle or a separation oracle to achieve sub-linear regret on manifolds. In the second part, we consider the last-iterate convergence of Riemannian extragradient-type methods, which can be provably employed to solve Riemannian minimax problems. We propose the Riemannian extragradient and the Riemannian past extragradient, demonstrating that both exhibit behavior analogous to that in Euclidean space.

Dr. Tao Luo (SJTU, Friday 11:00-11:45)

Bio: Dr. Tao Luo is an Associate Professor at the School of Mathematical Sciences/Institute of Natural Sciences at Shanghai Jiao Tong University (SJTU). His research focuses on the mathematical theory of machine learning and materials science. He received his Bachelor's degree in 2012 from the inaugural Science Class (now called Zhiyuan College) at SJTU and his Ph.D. from the Hong Kong University of Science and Technology (HKUST) in 2017, where he was awarded the Hong Kong Mathematical Society's Best PhD Thesis Award. From 2017 to 2020, he served as a Golomb Visiting Assistant Professor in the Department of Mathematics at Purdue University. He has made contributions to research in the frequency principle and condensation phenomena in deep learning, as well as the Peierls-Nabarro model, epitaxial growth, and the Cauchy-Born rule in materials science. His work has been published in top international conferences and journals, including the SIAM journal series, Arch. Ration. Mech. Anal., J. Mach. Learn. Res., NeurIPS, J. Mach. Learn., and CSIAM Trans. Appl. Math.

Title: The Theory of Parameter Condensation in Neural Networks

Abstract:

In this talk, we will first introduce the phenomenon of parameter condensation in neural networks, which refers to the tendency of certain parameters to converge towards the same values during training. Then, for certain types of networks, we prove that condensation occurs in the early stages of training. We further analyze which hyperparameters and training strategies influence parameter condensation. In some cases, we even provide a phase diagram that delineates whether parameter condensation occurs. We will also briefly discuss the relationship between parameter condensation and generalization ability. Finally, towards the end of the training, we study the set of global minima and present a detailed analysis of its geometric structure and convergence properties.

Dr. Zhichao Peng (HKUST, Friday 11:45-12:30)

Bio: Zhichao Peng is an assistant professor at the department of Mathematics, HKUST. Before joining HKUST in summer 2023, he was a postdoc at the department of Mathematics at Michigan State University. Zhichao obtained his doctoral degree from Rensselaer Polytechnic Institute in 2020. His main research interests include numerical methods for kinetic equations and wave equations, and data-driven dimensionality reduction techniques for numerical partial differential equations.

Title: A learning-based projection method for model order reduction of transport problems

Abstract:

Reduced order model (ROM) is a technique to reduce computational cost in numerical simulations for parametric problems or long-time simulations. Classical linear ROMs utilizes a low-dimensional linear space to approximate the underlying solution manifold. However, the Kolmogorov n-width, which is the optimal error of approximating the solution manifold with a linear space, may decay slowly as the space dimension grows for transport dominant problems. As a result, classical linear ROM may become inefficient or even inaccurate for such problems. Fortunately, intrinsic low-rank structures of transport problems may still be captured by leveraging some nonlinear transformations. Utilizing this fact, we propose to design a new learning-based ROM for transport problems.