Annual Summer School on Mathematical Aspects of Data Science
Research Talks' Abstracts

29 Jun–01 Jul 2026

June 4, 2026

Weighted Sampling for Online Causal Discovery

Arnab Bhattacharrya
University of Warwick, UK

Discovering the underlying causal structure of a system is a fundamental challenge in machine learning, requiring interventional data to distinguish between observationally equivalent models. We will discuss the problem of learning causal Bayesian networks in an online setting, where a learner sequentially observes individual samples from a stream of unknown interventional and observational distributions. We formulate this task as an online sequence prediction problem. To enable computational efficiency, we will show how dynamic programming algorithms originally developed for uniform DAG counting and sampling within Markov Equivalence Classes (MECs) can be extended to support score-decomposable, weighted sampling. Describes joint work with Philips George John, Sayantan Sen, and Vidya Sagar Sharma.

Sampling

Sinho Chewi
Yale University, USA

The tutorial will cover the following parts: (1) a basic introduction to the Langevin diffusion and its discretization analysis; (2) Metropolis–Hastings correction and the analysis of the Metropolis-adjusted Langevin algorithm (MALA); (3) the proximal sampler algorithm. Then, I will discuss how these ideas can be combined to produce a proximal gradient algorithm for composite log-concave sampling, following joint work with Linghai Liu.

Is Adaptive Sampling for Learning Worth It? A Minimax View

Kevin Jamieson
University of Washington, USA

Consider a set of possible actions, each with an unknown objective value that depends on the action through some function. By repeatedly selecting actions and observing noisy feedback, how many samples does it take to identify the best one — and does it help to choose those actions adaptively, using what we have already seen? We study this question through the tractable and revealing lens of linear function classes. The first part of the talk is a self-contained tutorial, assuming only linear algebra and basic probability: we trace the sample complexity of linear estimation from passive least-squares regression, through active learning and optimal experimental design (leverage scores, G-optimal designs), to identifying the best action. We build the key toolkit along the way — Gaussian width, matrix concentration, and information-theoretic lower bounds — framing everything through a minimax lens: worst-case over the unknown linear parameter but for a particular set of actions, with matching upper and lower bounds. This sets up a central puzzle: in the simplest such setting, adaptivity improves over the best fixed, non-adaptive design by only a logarithmic factor. Is it always this weak? The second part presents recent work, accepted at COLT 2026, answering this. We pin down the minimax sample complexity of non-adaptive designs via a Gaussian-width term, show that for many natural action sets adaptivity again buys only logarithmic factors, and then construct a set where adaptive sampling beats every non-adaptive design by a polynomial factor — the first separation of its kind — powered by a clean ℓ_2-norm estimation primitive. We close with what this reveals about when the geometry of the action set makes adaptivity pay off.

Diffusion Models

Holden Lee

Johns Hopkins University, USA

Diffusion models are a highly successful approach to generative modeling based on learning the score function (gradient of log-pdf) or similar proxy and then using it to simulate a Markov process (e.g., a stochastic differential equation) that transforms white noise into the data distribution. I will discuss two key theoretical questions for diffusion models: (1) Given a score function estimate, how efficiently and accurately can they generate a sample? (2) For what families of distributions can the score function be efficiently learned?

All-Purpose Mean Estimation over the Reals

Jasper Lee
University of California, Davis, USA

Even for statistical problems as fundamental as mean estimation, there is a theory-practice divide. Conventional methods like the sample mean, while supported by theoretical results under strong assumptions, are often brittle in the presence of extreme data. Practitioners thus often use ad-hoc and unprincipled "outlier removal" heuristics, revealing a marked theory-practice gap even for this very basic problem.

In this talk, I'll present a sharp 1-d mean estimator, whose error is optimal even in the leading multiplicative constant, under a minimal finite unknown variance assumption. I'll also discuss recent results on the robustness properties of this estimator, including its performance under infinite variance data and under adversarially corrupted data.

In terms of techniques, I will present a novel way to prove concentration bounds, via the perspective of mathematical programming and duality principles.

Time permitting, I will briefly discuss what is known for multivariate mean estimation with sharp constant guarantees, for "very high" dimensions and a forthcoming work for low dimensions.

Beyond Standard Deep Learning: New Frontiers in Nonparametric Estimation, Domain Generalization, and Transformer Theory

Yuanyuan Lin

Chinese University of Hong Kong, Hong Kong SAR

While deep learning has achieved remarkable empirical success, establishing rigorous theoretical guarantees and adaptive architectures for complex statistical tasks remains a crucial frontier. This talk presents a series of recent advancements that bridge the gap between theory and practice in nonparametric estimation, generative learning, and domain generalization using deep neural networks. We will explore four interconnected frameworks designed to overcome classical statistical bottlenecks:

- Data-Augmented Density Estimation: A novel nonparametric noise contrastive estimation (NCE) method that achieves simulation-free, asymptotically normalized density estimation with minimax optimal convergence rates, while adapting inherently to low-dimensional data structures.
- Unified Generative Regression: A joint framework for nonparametric regression and conditional distribution learning that utilizes a constrained deep generative model to handle multivariate outcomes and construct robust prediction intervals.
- Heterogeneity-Aware Domain Generalization: A domain-specific regression approach that leverages linear functionals of marginal source distributions and neural networks to mitigate the curse of dimensionality and ensure cross-domain predictive consistency.
- Theoretical Foundations of Transformers: A rigorous analysis of attention-only Transformers handling infinite-dimensional sequential inputs in nonparametric regression. By introducing the Generalized Event-stream Modulation Space (GEMS), we prove that Transformers can mitigate the curse of dimensionality and achieve optimal excess risk bounds without relying on standard feedforward layers.

These works provide new methodologies and non-asymptotic theoretical guarantees, demonstrating how tailored deep learning architectures can efficiently solve high-dimensional, heterogeneous, and sequential statistical problems.

Spectral Sparsification: from Graphs to Convex Cones

James Saunderson

Monash University, Australia

Spectral sparsification of graphs involves approximating a weighted graph with a sparse weighted graph such that the weighted Laplacians of the two graphs are spectrally close. This notion of approximation implies that many properties of the original graph are approximately preserved by its sparsified version. Remarkably, any weighted graph with n vertices has an epsilon-spectral sparsifier with $O(n/\epsilon^2)$ weighted edges, a celebrated result of Batson, Spielman, and Srivastava.

Notions of sparsification extend beyond graphs to other mathematical objects, ranging from the combinatorial (e.g., hypergraphs, linear codes) to the convex-geometric. In this talk I will give an introduction to spectral sparsification, before discussing recent work on one convex-geometric extension of the spectral sparsification model. In this extension, the object to be sparsified is a sum of points taken from a closed convex cone, and the aim is to approximate the sum (in a suitable sense with respect to the cone) by keeping only a small number of terms in the sum and suitably reweighting them. I will discuss extensions of certain results in spectral sparsification to this context, as well as a range of sparsification questions that remain open problems in this setting.

Foundations of Reinforcement Learning and Control

Claire Vernade

University of Technology Nuremberg, Germany

Reinforcement learning and control theory are two adjacent scientific fields that focus on optimizing the controller of unknown dynamical systems using feedback. While both fields have common roots in dynamic programming, they have evolved with distinct methodologies, goals, and cultures. Despite decades of mutual influence, a significant gap persists between the two communities. This tutorial introduces adaptive control, actor-critic reinforcement learning algorithms, and a new original way to combine these two paradigms for data-driven decision making on a classical locomotion control problem. Our aim is to provide keys to understand the core differences between both approaches, and insights to help experts and newcomers in each field better understand and engage with the tools and approaches of the other.

New Matrix Perturbation Bounds via Relative Norm

Van Vu

University of Hongkong, Hong Kong SAR

Matrix-perturbation bounds quantify how the spectral characteristics of a baseline (truth) matrix A change under additive noise E. Classical results, including Weyl's inequality for eigenvalues and the Davis–Kahan theorem for eigenvectors and eigenspaces, have long played a foundational role in mathematics. These bounds are known to be sharp in worst-case analysis.

In the 10 years, we have been working to develop a perturbation framework that leverages the interaction between E and the eigenvectors of A, leading to the notion of relative norm, which can be used to replace the operator norm of E in many applications. This perspective yields quantitative improvements over classical bounds, particularly when E is random, a common scenario in applications. In this case, one typically obtains a saving of order $n^{1/2}$, where n is the dimension.

This talk surveys these developments and main ideas, focusing on recent results concerning eigenspace perturbations. If time allows, we will discuss extensions to other spectral functionals and applications in different areas. (joint work with Phuc Tran, Vin University)

Functional Estimation

Pengkun Yang
Tsinghua University, China

Functional estimation asks for a low-dimensional property of a distribution or signal without necessarily learning the entire underlying object. This talk will give a mini-tutorial on polynomial methods that have become central to this area. Starting from functional estimation over large alphabets, I will explain why classical plug-in estimators fail, how best polynomial approximations lead to minimax-rate estimators, and how moment matching provides a complementary viewpoint for the fundamental limits. The second part of the talk will discuss recent developments that go beyond explicitly constructed polynomial estimators. I will introduce a nonparametric likelihood approach for unlabeled histograms, where the multiset of frequency counts is modeled through a mixture distribution and the NPMLE gives flexible plug-in estimators for functionals. I will also discuss minimax estimation with correlated observations, highlighting how dependence changes the effective difficulty of functional estimation.

Random Matrices and Deep Neural Networks

Pierre Youssef

New York University, Abu Dhabi, UAE

Deep neural networks are trained through backpropagation, and their trainability is closely tied to whether gradients explode or vanish. For fully connected ReLU networks, this boils down to the study of the spectral norm of a product of random matrices: the network Jacobian.

This talk will begin with a tutorial on the random matrix theory needed: i.i.d. random matrices, singular values and spectral norms, bulk behavior versus outliers, sparse and inhomogeneous random matrices, and the role of strong convergence and free probability in analyzing products of large random matrices. We will then apply these ideas to neural-network Jacobians and exploit recent breakthrough in the study of inhomogeneous random matrices and the strong convergence phenomenon. After recalling the classical i.i.d. initialization theory, we will derive a stability theorem for more structured models, including sparse networks arising from pruning and networks with weakly correlated weights. The theorem shows that, after the right normalization, these structured Jacobians have the same limiting spectral norm as the i.i.d initialized model.