Mathematics of Machine Learning
Les mathématiques de l'apprentissage machine
(Org: **Simone Brugiapaglia** (Concordia) and/et **Tim Hoheisel** (McGill))

NICHOLAS BARNFIELD, Harvard

MEHDI DAGDOUG, McGill University
Agnostic Model-Assisted Estimation in Finite Population Sampling

The aim of survey sampling is to estimate finite population parameters. In many settings, auxiliary information is available at the population level and can be incorporated through modeling within the model-assisted framework. Establishing the asymptotic properties of such estimators has traditionally been carried out on a method-by-method basis and has remained an active area of research over the past 25 years. Although these estimators often share similar properties, the techniques used to analyze them are typically tailored to the specific prediction method. A general theoretical framework applicable to arbitrary prediction methods is currently lacking.

In this work, we address this gap by introducing a novel form of conditional cross-fitting adapted to the dependence arising in survey sampling. Leveraging this approach, we develop a unified theory for the asymptotic behavior of model-assisted estimators based on general prediction methods. In particular, we establish the asymptotic normality of the proposed estimators and the consistency of a corresponding cross-fitting variance estimator, thereby enabling the construction of asymptotically valid confidence intervals. The proposed approach is shown to be asymptotically optimal in that it reaches the Godambe-Joshi lower bound. Simulation studies further demonstrate the favorable finite-sample performance of the proposed methodology.

ARIEL GOODWIN, Cornell University
Minimum enclosing ball algorithms in geodesic spaces

Finding the smallest ball enclosing a bounded set of points is a fundamental optimization problem in computational geometry and operations research, where it is sometimes known as the minimax facility location problem. Algorithms for solving this problem in Euclidean space are plentiful, but there are fewer viable options in non-Euclidean spaces like Riemannian manifolds and geodesic metric spaces. In this talk we highlight the role that geodesics have played in the design of minimum enclosing ball algorithms on non-Euclidean spaces, dating back to a finitely-convergent algorithm on tree graphs. We demonstrate that a simple greedy geodesic-following algorithm, known to converge in the Riemannian case, admits a simple complexity analysis in geodesic metric spaces satisfying an upper curvature bound. In the case of nonpositive curvature, we also propose a projection-based algorithm that generates certifiable lower bounds on the minimum enclosing ball radius.

MAROUANE IL IDRISSI, Université du Québec à Montréal
Generalized Hoeffding decomposition and the linear nature of non-linearities

Hoeffding's functional decomposition plays a crucial role in global sensitivity analysis and explainable AI, enabling the extraction of meaningful insights from black-box models. However, its application traditionally relies on the assumption of mutual independence among the model's inputs. Generalizing this result to non-mutually independent inputs has been an important challenge in the field of uncertainty quantification. In this talk, we will explore how this decomposition can be extended to accommodate dependent inputs, given two reasonable assumptions: non-perfect functional dependence and non-degenerate stochastic dependence. This generalization requires approaching the problem using a framework at the cornerstone of probability theory, functional analysis, and combinatorics. This generalization can be approached as a direct-sum decomposition problem of generated Lebesgue space, unveiling a surprisingly linear approach to handling stochastic and functional non-linearities. The proposed "ortho-canonical decomposition" relies on oblique projections rather than the traditional conditional expectations.

SOPHIE MORIN, Polytechnique Montréal

RACHEL MORRIS, Concordia University
Performance guarantees for deep Koopman autoencoders

From fluid flows to chaotic systems, the Koopman operator has emerged as a powerful tool to analyze dynamical systems and forecast data. Modern data-driven methods such as extended Dynamic Mode Decomposition provide a Galerkin approximation of the Koopman operator on a finite dictionary of user-defined functions. However since the quality of the approximation depends greatly upon the dictionary choice, a more recent approach to approximating the Koopman operator trains a deep autoencoder neural network to learn the basis functions from data. Universal approximation theorems justify the use of deep neural networks (DNNs) to learn the dictionary, but the network architecture, training plan, and amount of data required for high-quality approximations all merit further investigation. In this presentation, I will provide rigorous performance guarantees for learning the Koopman operator by DNNs via a sparsity-promoting training program. In particular, I will give insight into the accuracy and sample complexity of neural network Koopman approximations by providing generalization bounds based on recent function approximation techniques. Furthermore, I will highlight a DNN architecture and training routine that achieves these theoretical benchmarks in practice.

JANOSCH ORTMANN, UQAM
Para-Gaussian Approximations of Random Deep Neural Networks

It is known that randomly initialised neural networks have approximate Gaussian behaviour in the infinite-width limit. Unfortunately, this approximation can be poor for finite width and gets worse as one goes deeper into the network. In this talk, I will introduce a more accurate approximation that is slightly non-Gaussian. Constructed from the Hermite polynomials, we call it the para-Gaussian distribution. I will discuss how this para-Gaussian approximation can lead to a practical recursive method to approximate the deeper layers of the network. I will also explain how to apply this theory to linear neural networks (where things can be solved exactly) and to ReLU neural networks (where things can be approximated).

STEFANIA PETRA, Heidelberg University
Optimization Meets Information Geometry: Applications to Image Labeling

In the first part of the talk, I will introduce three complementary viewpoints on the local classification of metric data: a variational approach, a Markov random field (MRF) approach, and a dynamical systems approach based on information geometry. Starting from the first two viewpoints, I will outline how local decisions are coupled into non-local interacting decisions through regularization, and briefly discuss limitations with regard to learning model parameters from data.

In the second part, I will discuss the design of a dynamical system on an elementary statistical manifold, called the assignment flow, for the contextual classification of metric data given on a graph. I will then discuss learning its parameters from data using a linearization of the assignment flow and low-rank exponential integration. Finally, I will briefly present a geometric multilevel optimization approach to the labeling problem in a variational formulation, exploiting a natural hierarchy of models at different discretization levels.

KUNLUN QI, Michigan State University
Machine learning-assisted moment closure model for the linear Boltzmann equation

The Boltzmann equation, a fundamental equation in kinetic theory, serves as a bridge between microscopic particle dynamics and macroscopic continuum mechanics. However, deriving closed macroscopic moment systems from the Boltzmann equation remains a long-standing challenge due to the intrinsic non-closure of the moment hierarchy. In this talk, we introduce a machine learning (ML)-assisted moment closure model for the linear Boltzmann equation, addressing both the deterministic

and stochastic settings. Our approach leverages neural networks to learn the spatial gradient of the unclosed highest-order moment, enabling effective training through natural output normalization. For the deterministic problem, to ensure global hyperbolicity and stability, we derive and apply the constraints that enforce symmetrizable hyperbolicity of the system. For the stochastic problem, we adopt the generalized polynomial chaos (gPC)-based stochastic Galerkin method to discretize the random variables, resulting in a system for which the approach in the deterministic case can be used similarly. Several numerical experiments are shown to demonstrate the effectiveness and accuracy of our ML-assisted moment closure model for the linear Boltzmann equation with or without uncertainties.

GABRIEL RIOUX, Purdue University
Comparing Heterogeneous Data: an Optimal Transport Approach

Modern machine learning pipelines must handle heterogeneous datasets that differ in modality (e.g., images and text) or representation (e.g., when no canonical representation exists or representations vary across sources). Optimal transport (OT) provides a powerful geometric framework for comparing and transforming datasets, but requires a cost function that captures the desired correspondence between data points. For complex heterogeneous data, this cost is often chosen ad hoc and may fail to capture meaningful similarities.

Gromov–Wasserstein (GW) problems provide a quadratic generalization of OT that can naturally handle distributions on different spaces and align datasets, such as graphs, that lack a canonical representation. However, GW generally leads to a nonconvex quadratic optimization problem, complicating its statistical and computational analysis.

This talk will describe recent progress on the structural, statistical, and computational analysis of GW. The key result enabling these developments is a new representation of GW as the optimal value of an optimization problem whose objective depends on the input distributions only through a parameterized OT problem.

MIN WANG, Houston

XUEMENG WANG, Simon Fraser University
Into the danger zone: stable extrapolation in high-dimensional function and operator learning

Out-of-distribution (OOD) generalization is a central challenge in scientific machine learning (SciML). We study regression problems in which the test distribution differs from the training distribution and ask: under what assumptions on the target function or operator is stable extrapolation possible, and how far beyond the training domain can one extrapolate? Existing theory often controls the test error through additive penalties measuring the discrepancy between the training and test distributions. Such guarantees can be considerably more pessimistic than the generalization observed empirically.

Under holomorphic regularity assumptions on the target, we identify classes of functions and operators for which the OOD generalization error converges at algebraic rates. This phenomenon stems from the increasing smoothness of higher-index coordinates, leading to what we term a 'blessing of high dimensionality' We derive explicit rates for arbitrary test measures supported on suitable domains and quantify how the admissible domain depends on the underlying regularity. Our extrapolation guarantees are independent of the test distribution, depending only on its support. We establish matching rates for polynomial approximation, deep neural networks, and deep neural operators. In addition, we present extensive numerical experiments across a range of functions and operators that corroborate the predicted rates.

LAURA WEIDENSAGER, Simon Fraser University
Total Sensitivity Kernels: Learning Weighted ANOVA Kernels by RKHS Norm Minimization

Kernel methods with standard isotropic kernels treat all input variables similarly and therefore suffer from the curse of dimensionality. Many high-dimensional functions arising in applications have low effective dimension: their ANOVA (analysis of variance) decomposition is dominated by a few inputs and by main effects and low-order interactions. We introduce total

sensitivity kernels (TSKs), a family of weighted ANOVA kernels in which the weights of all ANOVA components are determined by a single sensitivity parameter per input.

All TSKs generate the same function space and differ only in their norms. Since the RKHS norm of the target function governs the approximation term in standard error bounds for kernel ridge regression, we choose the sensitivity parameters by minimizing this norm. We show that the resulting objective is strongly convex. The data-driven approximation, based on minimum-norm interpolation, converges uniformly on compact parameter sets as the number of samples grows. Numerical experiments show substantial improvements over product kernels and unweighted ANOVA kernels, and illustrate that TSKs exploit multivariable structures.

This talk is based on joint work with John Darges (Emory University).

ANDRE WIBISONO, Yale University

Hamiltonian Dynamics for Accelerated Optimization and Sampling

Optimization and sampling are fundamental computational tasks in machine learning. Classical algorithms are based on discretizations of greedy dynamics, such as gradient flow for optimization and Langevin dynamics for sampling. In this talk, we explore how Hamiltonian dynamics provides a common framework for acceleration. Concretely, simulating Hamiltonian dynamics while periodically resetting the velocity to zero gives a descent method for optimization, while refreshing the velocity from a Gaussian distribution gives Hamiltonian Monte Carlo for sampling. We show that short integration times lead to nonaccelerated rates analogous to greedy methods. With longer integration times and suitable averaging, we obtain accelerated rates for optimization and sampling. Both analyses rely on the same averaging property of Hamiltonian flow.