Mathematics of Machine Learning
Mathématiques de l'apprentissage automatique

(Org: **Ben Adcock** (Simon Fraser University), **Ricardo Baptista** (University of Toronto) and/et **Giang Tran** (University of Waterloo))

HUNG-HSU CHOU, University of Pittsburgh
More is Less: Understanding Compressibility of Neural Networks via Implicit Regularization and Neural Collapse

Despite their recent successes, most modern machine learning algorithms lack theoretical guarantees, which are crucial to further development towards delicate tasks. One mysterious phenomenon is that, among infinitely many possible ways to fit data, the algorithms often find the "good" ones, even when the definition of "good" is not specified by the designers. In this talk I will approach this from both the microscopic view and the macroscopic view, with empirical and theoretical study of the connection between the good solutions in neural networks and the sparse solutions in compressed sensing. The key concepts are the implicit bias/regularization in machine learning models, and the neural collapse phenomenon induced by the block structure of neural tangent kernel, which can be used for out-of-distribution detection.

ISAAC GIBBS, University of California, Berkeley
Designing probabilistic predictors for multiple decision makers

We consider the problem of constructing probabilistic predictions that lead to accurate decisions when employed by downstream users to inform actions. For a single decision maker, designing an optimal predictor is equivalent to minimizing a proper loss function corresponding to the negative utility of that individual. For multiple decision makers, our problem can be viewed as a variant of omniprediction in which the goal is to design a single predictor that simultaneously minimizes multiple losses. We will discuss two strategies for designing sample-efficient algorithms for this problem. The first is a two-player game based approach in which the two players alternate between estimating and responding to the worst-case loss. The second is a more direct procedure that exploits structural properties of the set of proper losses. Empirical evaluations show that both of these methods perform well in practice.

AVI GUPTA, Simon Fraser University
Universal Nonlinear Learning of High-Dimensional Anisotropic Sobolev Functions from Point Samples

A central problem in scientific machine learning is high-dimensional function recovery from limited data. Widths, an important concept from information-based complexity, provide a standard way to quantify this. At the same time, universal approximation theorems highlight the representational power of nonlinear models such as neural networks and have become a central theme in machine learning. In anisotropic settings, where different coordinates exhibit different smoothness, universality questions arise naturally. In this work, we study them for recovery from point samples (standard information) in periodic Sobolev spaces with anisotropic smoothness, including both anisotropic Sobolev and anisotropic mixed-smoothness Sobolev spaces. Our approach is a theoretical nonlinear reconstruction scheme inspired by compressed sensing, for which we derive worst-case upper bounds on the recovery error. We prove a universal approximation result by showing that our nonlinear reconstruction map achieves asymptotically better error guarantees, simultaneously over a family of anisotropic smoothness classes, than any linear method, thus justifying nonlinear algorithms for universal approximation in such spaces. In terms of widths, we establish matching upper and lower bounds for nonlinear widths, thereby identifying the optimal performance of sampling-based recovery on these spaces. Such results on nonlinear widths in anisotropic settings have been largely absent, so our bounds are of concrete interest for sampling recovery and their learning-theoretic applications.

MOHAMED HIBAT-ALLAH, University of Waterloo
Language models for quantum many-body physics

Recent large language models have achieved remarkable success, performing at near-human levels on many tasks such as speech recognition, machine translation, and text generation. In this talk, I will show how we can adapt language-model architectures, particularly recurrent neural networks (RNNs), to study quantum many-body systems. By training these models on quantum many-body physics problems, we obtain results that are competitive with traditional numerical quantum simulation methods. This progress highlights the exciting possibility of bringing insights from modern language models to quantum simulation.

SPENCER HILL, Queen's University
Communication Complexity of Exact Sampling under Exponential Cost

Exact sampling is the problem of communicating a message so that, given a shared source of randomness, a receiver can generate a random sample with a prescribed probability distribution. This problem arises naturally in many forms of learned data compression, most notably as a promising alternative to quantization in nonlinear transform coding.

In this talk, I will describe some applications of exact sampling in machine learning and several state-of-the-art sampling algorithms. I will provide an overview of recent work in the exponential cost setting, presenting matching upper and lower bounds and discussing surprising differences in algorithm performance in this generalized setup.

SHIKHAR JAISWAL, University of Toronto
Understanding The Modality Gap In Multi-Modal Systems

Multi-modal systems are integral components of machine learning models that process information from various data modalities, like text, images and audio by mapping these inputs into a shared representation space. Inherent to the development of these systems is the phenomenon of "modality gap", wherein, image embeddings and text embeddings are "distributed" differently within the shared representation space. This "gap" between the learnt representations has detrimental consequences for downstream tasks like search, retrieval and recommendation. While several empirical studies have attributed this phenomenon to the inherent difference in the distributional richness of the modalities, the mathematical analysis of its origins remains limited. In this talk, we will (i) define the modality gap in a generalized way, with a focus on its effect for downstream applications; (ii) present recent studies to better understand and mitigate this issue; and (iii) analyze the problem under simplifying assumptions regarding the network architecture and data distribution.

ANASTASIS KRATSIOS, McMaster University
Incremental Generation is Necessity and Sufficient for Universality in Flow-Based Modelling

Incremental flow-based denoising models have reshaped generative modelling, but their empirical advantage still lacks a rigorous approximation-theoretic foundation. We show that incremental generation is necessary and sufficient for universal flow-based generation on the largest natural class of self-maps of $[0,1]^d$ compatible with denoising pipelines, namely the orientation-preserving homeomorphisms of $[0,1]^d$. All our guarantees are uniform on the underlying maps and hence imply approximation both samplewise and in distribution.

Using a new topological-dynamical argument, we first prove an impossibility theorem: the class of all single-step autonomous flows, independently of the architecture, width, depth, or Lipschitz activation of the underlying neural network, is meagre and therefore not universal in the space of orientation-preserving homeomorphisms of $[0,1]^d$. By exploiting algebraic properties of autonomous flows, we conversely show that every orientation-preserving Lipschitz homeomorphism on $[0,1]^d$ can be approximated at rate $\mathcal{O}(n^{-1/d})$ by a composition of at most K_d such flows, where K_d depends only on the dimension. Under additional smoothness assumptions, the approximation rate can be made dimension-free, and K_d can be chosen uniformly over the class being approximated. Finally, by linearly lifting the domain into one higher dimension, we obtain structured universal approximation results for continuous functions and for probability measures on $[0,1]^d$, the latter realized as pushforwards of empirical measures with vanishing 1-Wasserstein error.

Joint work: Hossein Rouhvarzi

SOPHIE MORIN, Polytechnique Montréal
Equivariant machine learning for collision detection of ellipses and related shapes

Computing the distance between two objects in space, in particular determining whether they have collided or are very close to doing so, is essential in a wide range of computational applications. It is trivial when the objects are spheres, line segments, or other very simple shapes, and good algorithms are known for polytopes. However, something as apparently simple as the distance between two ellipses in the plane remains surprisingly difficult if one wants both speed and accuracy. In this talk, I will discuss an equivariant machine learning framework for this problem and present some results from ongoing work.

RACHEL MORRIS, Concordia University
Regularity guarantees for adversarially robust learning

While neural network image classification enjoys high success rates in most settings, recent work discovered that well-targeted adversarial attacks can transform a correctly classified image into one that is visually indistinguishable from the original but that completely fools the classification algorithm. This has sparked many new approaches to classification which include an adversary in the training process: such an adversary can improve robustness and generalization properties, at the cost of decreased accuracy and increased training time. By considering a "worst-case" adversary, the resulting mathematical model for adversarial training can be understood as an energy minimization problem with a regularizing nonlocal perimeter term. In this presentation, I will discuss my current work studying regularity guarantees for the decision boundary of an adversarially robust minimizer. In particular, for a continuous and bounded underlying density, the decision boundary is C^2 smooth. I will discuss using explicit geometric perturbations and second variation analysis to show singular points (i.e. corners, cusps) are suboptimal. For the smoother points, I will demonstrate how leveraging necessary conditions allows one to upgrade C^1 regularity to C^2 regularity.

CAMERON MUSCO, University of Massachusetts Amherst
Structured Matrix Approximation via Matrix-Vector Products

In this talk, I will give an overview of recent progress on the problem of structured matrix approximation from matrix-vector products. Given a target matrix A that can only be accessed through a limited number of (possibly adaptively chosen) matrix-vector products, we seek to find a near-optimal approximation to A from some structured matrix class – e.g., a low-rank approximation, a hierarchical low-rank approximation, a sparse or diagonal approximation, etc. This general problem arises across the computational sciences and data science, both in algorithmic applications and, more recently, in scientific machine learning, where it is closely related to the problem of linear operator learning from input/output samples.

I will overview recent work, where we give 1) optimal algorithms for approximating A with a matrix with a fixed sparsity pattern (e.g., a diagonal or banded matrix), 2) the first algorithms with strong relative error bounds for hierarchical low-rank approximation, and 3) the first bounds for generic structured families with sample complexity depending on the parametric complexity of the family. I will highlight several open questions on structured matrix approximation and its applications to operator learning.

ESHA SAHA, University of Alberta
Data-Driven Solutions to Coupled PDEs using Disjoint Priors

Advances in data acquisition and computational power have led to a rapid increase in high-dimensional (ODE or PDE) modelling. In many applications, especially in biological and ecological modeling, the primary challenge is not data unavailability but the existence of data that is incomplete, making it either useless or the entire data collection effort a waste of resources. Complex phenomena are often described by coupled (or more) variables, yet only a subset is supported by known governing equations, while the remaining variables are available only through data. This mismatch between known physics and observed data creates difficulties for finding solutions to the model, even with the well-known physics-informed machine learning techniques since they typically assume full knowledge of either the system physics or complete data across all variables. In this presentation, I will discuss some ground challenges in modelling partially observed, coupled systems and demonstrate how a neural-network-based approach can effectively solve them even when the variables constrained by physics and those informed by data are mutually exclusive.

MATTHEW THORPE, University of Warwick
How Many Labels Do You Need in Semi-Supervised Learning?

Semi-supervised learning (SSL) is the problem of finding missing labels from a partially labelled data set. The heuristic one uses is that "similar feature vectors should have similar labels". The notion of similarity between feature vectors explored in this talk comes from a graph-based geometry where an edge is placed between feature vectors that are closer than some connectivity radius. A natural variational solution to the SSL is to minimise a Dirichlet energy built from the graph topology. And a natural question is to ask what happens as the number of feature vectors goes to infinity? In this talk I will give results on the asymptotics of graph-based SSL using an optimal transport topology. The results will include a lower bound on the number of labels needed for consistency.

ALEX TOWNSEND, Cornell University
A Mathematical Guide to Operator Learning

A fundamental challenge in modern scientific computing is learning an operator from finite data. In this talk, we offer a mathematical guide to operator learning, drawing a distinction between passive and active observation models and revealing the crucial role this choice plays in sample efficiency. We explore how the nature of the underlying partial differential equation, i.e., elliptic, parabolic, or hyperbolic, governs the difficulty of learning the associated solution operator, and we present recent learning theory that quantifies the number of queries needed for accurate recovery. Diffusive systems, as we shall see, are forgiving; wave-like systems are not. Along the way, we reflect on what it means to learn in infinite dimensions and how mathematical structure can be exploited to tame the curse of dimensionality.

YUNAN YANG, Cornell University
Training Distribution Optimization in the Space of Probability Measures

A central question in data-driven modeling is: from which probability distribution should training samples be drawn to most effectively approximate a target function or operator? This work addresses this question in the setting where "effectiveness" is measured by out-of-distribution (OOD) generalization accuracy across a family of downstream tasks. We formulate the problem as minimizing the expected OOD generalization error, or an upper bound thereof, over the space of probability measures. The optimal sampling distribution depends jointly on the model class (e.g., kernel regressors, neural networks), the evaluation metric, and the target map itself. Building on this characterization, we propose two adaptive, target-dependent data selection algorithms based on bilevel and alternating optimization. The resulting surrogate models exhibit significantly improved robustness to distributional shifts and consistently outperform models trained with conventional, non-adaptive, or target-independent sampling across benchmark problems in function approximation, operator learning, and inverse modeling.