Mathematics of Machine Learning

(Org: **Ben Adcock** (Simon Fraser University), **Elina Robeva** (UBC) and/et **Giang Tran** (University of Waterloo))

RICARDO BAPTISTA, California Institute of Technology

Dynamics and Memorization Behaviour of Score-Based Diffusion Models

Diffusion models have emerged as a powerful framework for generative modeling that relies on score matching to learn gradients of the data distribution's log-density. A key element for the success of diffusion models is that the optimal score function is not identified when solving the denoising score matching problem. In fact, the optimal score in both unconditioned and conditioned settings leads to a diffusion model that returns to the training samples and effectively memorizes the data distribution. In this presentation, we study the dynamical system associated with the optimal score and describe its long-term behavior relative to the training samples. Lastly, we show the effect of two forms of score function regularization on avoiding memorization: restricting the score's approximation space and early stopping of the training process. These results are numerically validated using distributions with and without densities including image-based problems.

BENJAMIN BLOEM-REDDY, University of British Columbia

Causal Inference with Cocycles

Many interventions in causal inference can be represented as transformations of the variables of interest. Abstracting interventions in this way allows us to identify a local symmetry property exhibited by many causal models under interventions. Where present, this symmetry can be characterized by a type of map called a cocycle, an object that is central to dynamical systems theory. We show that such cocycles exist under general conditions and are sufficient to identify interventional distributions and, under suitable assumptions, counterfactual distributions. We use these results to derive cocycle-based estimators for causal estimands and show that they achieve semiparametric efficiency under standard conditions. Since entire families of distributions can share the same cocycle, these estimators can make causal inference robust to mis-specification by sidestepping superfluous modelling assumptions. We demonstrate both robustness and state-of-the-art performance in several simulations, and apply our method to estimate the effects of 401(k) pension plan eligibility on asset accumulation using econometric data.

Based on joint work with Hugh Dance (UCL/Gatsby Unit): https://arxiv.org/abs/2405.13844

WUYANG CHEN, Simon Fraser University

Towards Data-Efficient and OOD Generalization of Scientific Machine Learning Models

In recent years, there has been growing promise in coupling machine learning methods with domain-specific physical insights to solve scientific problems based on partial differential equations (PDEs). However, there are two critical bottlenecks that must be addressed before scientific machine learning (SciML) can become practically useful. First, SciML requires extensive pretraining data to cover diverse physical systems and real-world scenarios. Second, SciML models often perform poorly when confronted with unseen data distributions that deviate from the training source, even when dealing with samples from the same physical systems that have only slight differences in physical parameters. In this line of work, we aim to address these challenges using data-centric approaches. To enhance data efficiency, we have developed the first unsupervised learning method for neural operators. Our approach involves mining unlabeled PDE data without relying on heavy numerical simulations. We demonstrate that unsupervised pretraining can consistently reduce the number of simulated samples required during fine-tuning across a wide range of PDEs and real-world problems. Furthermore, to evaluate and improve the out-of-distribution (OOD) generalization of neural operators, we have carefully designed a benchmark that includes diverse physical parameters to emulate real-world scenarios. By evaluating popular architectures across a broad spectrum of PDEs, we conclude that neural operators achieve more robust OOD generalization when pretrained on physical dynamics with high-frequency patterns rather than smooth ones. This suggests that data-driven SciML methods will benefit more from learning from challenging samples.

HANS DE STERCK, University of Waterloo
Fast Multipole Attention for Transformer Neural Networks

Transformer-based machine learning models have achieved state-of-the-art performance in many areas. However, the quadratic complexity of the self-attention mechanism in Transformer models with respect to the input length hinders the applicability of Transformer-based models to long sequences. To address this, we present Fast Multipole Attention (FMA), a new attention mechanism that uses a divide-and-conquer strategy to reduce the time and memory complexity of attention for sequences of length n from $\mathcal{O}(n^2)$ to $\mathcal{O}(n \log n)$ or $\mathcal{O}(n)$, while retaining a global receptive field. The hierarchical approach groups queries, keys, and values into $\mathcal{O}(\log n)$ levels of resolution, where groups at greater distances are increasingly larger in size and the weights to compute group quantities are learned. As such, the interaction between tokens far from each other is considered in lower resolution in an efficient hierarchical manner. This multi-level divide-and-conquer strategy is inspired by fast summation methods from n-body physics and the Fast Multipole Method. We perform evaluation on autoregressive and bidirectional language modeling tasks and compare our FMA model with other efficient attention variants on medium-size datasets. We find empirically that the Fast Multipole Transformer outperforms other efficient transformers in terms of memory size and accuracy. The FMA mechanism has the potential to empower large language models with greater sequence lengths, taking the full context into account in an efficient, naturally hierarchical manner during training and when generating long sequences.

NICK HARVEY, University of British Columbia
When Online Learning Meets Stochastic Calculus

Online learning is a theoretical framework for learning and optimization without statistical assumptions on the data. Optimization methods developed in this setting are usually robust and have formal notions of worst-case or adaptive performance. A recent line of work has looked at online learning through the lens of differential equations and continuous-time analysis. This viewpoint has yielded new understanding of classical results, and has also led to new optimal results for several problems. In this talk I will discuss a few uses of stochastic calculus in the design and analysis of online learning methods, focusing on the classical problem of prediction with experts' advice.

Joint work with Chris Liaw (Google Research), Sikander Randhawa (UBC) and Victor Sanches Portella (University of São Paulo).

MIRANDA HOLMES-CERFON, University of British Columbia
Programmable assembly: inverse design of materials from discrete components

Particles and discrete objects on the scale of nanometres to micrometres, such as colloids, DNA bricks, proteins, transistors, etc, are increasingly being used as building blocks for new materials, with a variety of applications such as in optics, drug delivery, energy harvesting, and nanorobotics. A goal for theory and simulation is to build algorithms and design principles to find the building blocks that assemble into a structure of interest. This is a challenge due to the high-dimensionality of the systems of interest, the presence of strong noise, and the sometimes far-from-equilibrium conditions, making standard optimization algorithms inapplicable and demanding new approaches. I will describe our group's progress on finding optimal conditions for "addressable self-assembly", a system of particles where each building block is distinct and has a specific location in the target structure, and that assembles spontaneously under thermal fluctuations. I will show how using tools derived from machine learning can generate novel solutions for small systems, and will point out challenges in extending these tools to more complex systems.

SAMUEL LANTHALER, Caltech
Generative AI for the statistical computation of fluids

In recent years, there has been growing interest in the use of neural networks for the data-driven approximation of PDE solution operators. This talk will focus on a recent application of neural networks to the statistical computation of fluid flows. In this application, the choice of training objective is observed to lead to stark differences in the empirically achieved results.

I will argue that implicit constraints, related to limitations of what is practically achievable by deep learning, could provide a theoretical explanation of these observations.

MATHIAS LECUYER, University of British Columbia
Adaptive Randomized Smoothing: Certified Adversarial Robustness for Multi-Step Defences

ML theory usually considers model behaviour in expectation. In practical deployments however, we often expect models to be robust to adversarial perturbations, in which a user applies deliberate changes to on input to influence the prediction a target model. For instance, such attacks have been used to jailbreak aligned foundation models out of their normal behaviour. Given the complex models that we now deploy, how can we enforce such robustness properties while keeping model flexibility and utility?

I will present recent work on Adaptive Randomized Smoothing (ARS), an approach we developed to certify the predictions of test-time adaptive models against adversarial examples. ARS extends the analysis of randomized smoothing using f-Differential Privacy, to certify the adaptive composition of multiple steps during model prediction. We show how to instantiate ARS on deep image classification to certify predictions against adversarial examples of bounded ℓ_∞ norm.

KE LI, Simon Fraser University
Rethinking Regression: Insights from Machine Learning

Regression problems arise every time one would like to predict a continuous-valued variable, be it the colour of a pixel, a 3D position, a system configuration or a feature vector. It is well known that regression with square loss yields the conditional mean as the prediction. This is undesirable when there could be many predictions that are all correct, since the conditional mean would effectively average over these predictions and could be far from any of them. As an example, when the prediction takes the form of an image, the conditional mean tends to be blurry and desaturated. On the other hand, in classification problems, ambiguity in labels does not cause an issue because classifiers produce a distribution over class labels as output. Is it possible to get the best of both worlds? In this talk, I will show how to do so using a simple technique, known as conditional Implicit Maximum Likelihood Estimation.

WENLONG MOU, University of Toronto
Continuous-time reinforcement learning: blessings of elliptic structures and high-order approximations

Reinforcement learning (RL) for controlling continuous-time diffusion processes has attracted significant research interest in recent years. A key challenge is accurately estimating the value function for an unknown system, given only discretely observed trajectory data. While model-free RL methods offer the flexibility of advanced function approximations, they struggle with long effective horizons and lack the precision of model-based approaches.

In this talk, I present recent developments in the design of continuous-time policy evaluation algorithms, introducing a novel class of Bellman equations. These methods integrate the flexibility of RL techniques with the precision of high-order numerical schemes. Among other results, I will highlight how the underlying elliptic structures provide strong theoretical guarantees, even as the effective horizon extends to infinity. Finally, I will discuss how these theoretical insights inform practical algorithmic design.

RAHUL PARHI, University of California, San Diego
Deep Learning Meets Sparse Regularization

Deep learning has been wildly successful in practice and most state-of-the-art artificial intelligence systems are based on neural networks. Lacking, however, is a rigorous mathematical theory that adequately explains the amazing performance of deep neural networks. In this talk, I present a new mathematical framework that provides the beginning of a deeper understanding of deep learning. This framework precisely characterizes the functional properties of trained neural networks. The key mathematical tools which support this framework include transform-domain sparse regularization, the Radon transform

of computed tomography, and approximation theory. This framework explains the effect of weight decay regularization in neural network training, the importance of skip connections and low-rank weight matrices in network architectures, the role of sparsity in neural networks, and explains why neural networks can perform well in high-dimensional problems.

DANICA SUTHERLAND, University of British Columbia
Expander Graphs and Low-Distortion Embeddings for Learning on Graphs

Graphs are a natural model for many domains we would like to learn on, and graph neural networks based on local message passing have seen success on various problems. In other areas of machine learning, however, Transformers (based on pairwise "attention") are the dominant recent machine learning model. Yet Graph Transformers have had significant scaling problems because of the quadratic full everything-to-everything attention. This talk presents a line of work addressing this problem: first, in Exphormer ("expander" + "transformer"), we exploit expander graphs to create a sparse graph to augment the original problem graph, for limited attention but good expansion properties across layers. Exphormer helps scale (computationally and statistically) graph Transformers to much larger graphs, obtaining state-of-the-art results on many kinds of graph problems. On many very large graphs (e.g. social networks or protein-protein interaction networks), though, even the original problem's graph is too large for practical learning; we thus build an extension called Spexphormer ("sparse Exphormer"), which further constrains attention to "important" edges, dramatically reducing memory usage. We finally present a theoretical account of situations where Spexphormer's sparsification is possible, and where it is not.

CHRISTOS THRAMPOULIDIS, University of British Columbia
Implicit Geometry of Next-token Prediction: From Language Sparsity Patterns to Model Representations

How do language models map linguistic patterns to their representations? Specifically, can the geometry of context and word embeddings be characterized by the structure of the training data? We demonstrate that, in a large-model regime trained sufficiently long, context and word embeddings emerge from matrix factorization of a logit matrix that decomposes into sparse and low-rank components. As training progresses, the low-rank component becomes dominant and can be computed solely from the sparsity pattern of the training data, determined by unique word-context pairings across the dataset.

SHARAN VASWANI, Simon Fraser University
Global Convergence of Softmax Policy Gradient for Stochastic Bandits

Though policy gradient (PG) methods have played a vital role in the achievements of reinforcement learning (RL), the theoretical understanding of these methods is quite limited. Consequently, we focus on stochastic bandit problems (the simplest RL setting) and study the convergence of softmax policy gradient (SPG), a commonly used RL algorithm. Despite the non-concavity of the underlying objective function, recent research has leveraged the objective's smoothness and gradient domination properties to establish the convergence of SPG to an optimal policy. However, these results require setting the SPG parameters according to unknown problem-dependent quantities (e.g. the optimal action or the true reward vector in a bandit problem). We address this limitation by proposing to use SPG with exponentially decreasing step sizes. Specifically, we prove that the resulting algorithm offers similar theoretical guarantees as the state-of-the-art without requiring the knowledge of oracle-like quantities. However, using such decreasing step-sizes adversely affects the algorithm's empirical performance. Consequently, we analyze the algorithm from a different perspective and show that SPG with any constant step-size can asymptotically converge to a globally optimal policy almost surely.

ANDREW WARREN, University of British Columbia
Estimation of one-dimensional structures from noisy empirical observation

Given a data distribution which is concentrated around a one-dimensional structure, can we infer that structure? We consider versions of this problem where the distribution resides in a metric space and the 1d structure is assumed to either be the range of an absolutely continuous curve, a connected set of finite 1d Hausdorff measure, or a general 1-rectifiable set. In each of

these cases, we relate the inference task to solving a variational problem where there is a tradeoff between data fidelity and simplicity of the inferred structure; the variational problems we consider are closely related to the so-called "principal curve" problem of Hastie and Steutzle as well as the "average-distance problem" of Buttazzo, Oudet, and Stepanov. For each of the variational problems under consideration, we establish existence of minimizers, stability with respect to the data distribution, and consistency of a discretization scheme which is amenable to Lloyd-type numerical methods. Lastly, we consider applications to estimation of stochastic processes from partial observation, as well as the lineage tracing problem from mathematical biology.

YIMING XU, University of Kentucky
Statistical Ranking with Dynamic Covariates

The Plackett-Luce model has been widely applied for rank aggregation in sports analytics and social sciences. In this presentation, we consider a covariate-assisted ranking model within the Plackett-Luce framework. Unlike existing approaches focusing solely on pure covariates or individual effects with fixed covariates, our model incorporates individual effects with dynamic covariates. This increased flexibility enhances model fitting by allowing for individualized dynamic ranking but also presents significant challenges in analysis. We address these challenges in the context of maximum likelihood estimation (MLE) under a general graph topology. Specifically, we provide conditions for model identifiability and the unique existence of the MLE, propose an alternating maximization algorithm to compute the MLE, and establish a uniform consistency result. Finally, we demonstrate an application of the proposed model by analyzing a large-scale ATP tennis dataset.

OZGUR YILMAZ, The University of British Columbia
Generative compressed sensing with Fourier measurements

In the recent years, it has been established that Deep Generative Models (DGMs) can be used as priors in inverse problems such as denoising, inpainting, medical and seismic imaging, and more. One inverse problem of tremendous interest since 2005 is compressed sensing (CS) – acquisition and provable recovery of sparse signals (or signals with low complexity) from a few, non-adaptive measurements. Recently DGMs have been proposed to replace the sparse signal model in CS, leading to theoretical guarantees and practical performance that improves on "classical compressed sensing" for classes of signals that can modelled well using DGMs when the measurement matrix and/or network weights follow a subgaussian distribution. We move beyond the subgaussian assumption, to measurement matrices that are derived by sampling rows of a unitary matrix (including subsampled Fourier measurements as a special case). Specifically, we construct model-adapted sampling strategies and prove restricted isometry guarantee for generative compressed sensing with subsampled isometries, leading to recovery bounds with nearly order-optimal sample complexity.