Mathematics of Machine Learning
Les mathématiques de l'apprentissage automatique
(Org: **Ben Adcock** (SFU), **Jason Bramburger** (Concordia), **Giang Tran** (Waterloo) and/et **Hamid Usefi** (Memorial))

ZIAD ALDIRANY, Polytechnique Montréal
Multi-Level Approach for Error Reduction in Physics-Informed Neural Networks

In recent years, deep learning approaches, such as the physics-informed neural networks (PINNs), have shown promising results for several classes of initial and boundary-value problems. However, their ability to surpass, particularly in terms of accuracy, classical discretization methods such as the finite element methods, remains a significant challenge. One of the main obstacles of deep learning approaches lies in their inability to consistently reduce the relative error in the computed solution. We present our novel approach, the multi-level neural networks, in order to reduce the solution error when using deep learning approaches. The main idea consists in computing an initial approximation to the problem using a simple neural network and in estimating, in an iterative manner, a correction by solving the problem for the residual error with a new network of increasing complexity. This sequential reduction of the residual associated with the partial differential equation allows one to decrease the solution error, which, in some cases, can be reduced to machine precision. The underlying explanation is that the method is able to capture at each level smaller scales of the solution using a new network. Numerical examples in 1D and 2D dealing with linear and non-linear problems are presented to demonstrate the effectiveness of the proposed approach using PINNs.

SIMONE BRUGIAPAGLIA, Concordia University
Generalization limits of deep neural networks in identity effects learning

A key open problem in the mathematical foundations of deep learning is understanding *generalization*, informally defined as the ability of neural networks to successfully perform a given task outside the training set. Motivated by this challenge and by applications to cognitive science, we consider the problem of learning *identity effects*, i.e., classifying whether a pair of objects is identical or not, and present a theory aimed at rigorously identifying the generalization limits of deep learning for this task.

First, we will illustrate a general *rating impossibility* theorem that identifies settings where machine learning algorithms are provably unable to generalize outside the training set. Then, we will show how to apply this theorem to popular deep learning architectures such as feed-forward, recurrent and graph neural networks trained via stochastic gradient descent or Adam. For graph neural networks, we will also present a *rating possibility* theorem that establishes sufficient conditions for the existence of architectures able to generalize outside the training set. Finally, we will illustrate numerical experiments that either validate our theoretical findings or identify gaps between theory and practice.

This presentation is based on joint work with Giuseppe A. D'Inverno, Matthew Liu, Mirco Ravanelli, and Paul Tupper.

ELIZABETH COLLINS-WOODFIN, McGill University
High dimensional limit of streaming SGD for generalized linear models

We provide a characterization of the high dimensional limit of one-pass, single batch stochastic gradient descent (SGD) in the case where the number of samples scales proportionally with the problem dimension. We characterize the limiting process in terms of its convergence to a high-dimensional stochastic differential equation, referred to as the homogenized SGD. Our proofs assume Gaussian data but allow for a very general covariance structure. Our set-up covers a range of optimization problems including linear regression, logistic regression, and some simple neural nets. For each of these models, the convergence of SGD to homogenized SGD enables us to derive a close approximation of the statistical risk (with explicit and vanishing error bounds) as the solution to a Volterra integral equation. In a separate paper, we perform similar analysis without the Gaussian assumption in the case of SGD for linear regression. (Based on joint work with C. Paquette, E. Paquette, I. Seroussi).

ADAM GARDNER, Artinus Consulting Inc.
Decoding Neural Scaling Laws

For a large variety of models and datasets, neural network performance has been empirically observed to scale as a power-law with model size and dataset size. We will explore the origins of these scaling laws and their relationship to geometric proprieties such as the dimension of the data manifold and symmetries shared by the model and dataset. While the takeaway from these scaling laws for many prominent artificial intelligence labs is to improve performance by increasing model and dataset sizes, we propose an alternative perspective - a deeper mathematical understanding of these scaling laws will help researchers discover more efficient neural network architectures. We conclude with some potential future directions for this line of research.

MARK IWEN, Michigan State University
Sparse Spectral Methods for Solving High-Dimensional and Multiscale Elliptic PDEs

In his monograph "Chebyshev and Fourier Spectral Methods", John Boyd claimed that, regarding Fourier spectral methods for solving differential equations, "[t]he virtues of the Fast Fourier Transform will continue to improve as the relentless march to larger and larger [bandwidths] continues". This talk will discuss attempts to further the virtue of the Fast Fourier Transform (FFT) as not only bandwidth is pushed to its limits, but also the dimension of the problem. Instead of using the traditional FFT however, we make a key substitution from the sublinear-time compressive sensing literature: a high-dimensional, sparse Fourier transform (SFT) paired with randomized rank-1 lattice methods. The resulting sparse spectral method rapidly and automatically determines a set of Fourier basis functions whose span is guaranteed to contain an accurate approximation of the solution of a given elliptic PDE. This much smaller, near-optimal Fourier basis is then used to efficiently solve the given PDE in a runtime which only depends on the PDE's data/solution compressibility and ellipticity properties, while breaking the curse of dimensionality and relieving linear dependence on any multiscale structure in the original problem. Theoretical performance of the method is established with convergence analysis in the Sobolev norm for a general class of nonconstant diffusion equations, as well as pointers to technical extensions of the convergence analysis to more general advection-diffusion-reaction equations. Numerical experiments demonstrate good empirical performance on several multiscale and high-dimensional example problems, further showcasing the promise of the proposed methods in practice.

WENJING LIAO, Georgia Institute of Technology
Exploiting low-dimensional structures in machine learning and PDE simulations

Many data in real-world applications are in a high-dimensional space but exhibit low-dimensional structures. In mathematics, these data can be modeled as random samples on a low-dimensional manifold. I will talk about machine learning tasks like regression and classification, as well as PDE simulations. We consider deep learning as a tool to solve these problems. When data are sampled on a low-dimensional manifold, the sample complexity crucially depends on the intrinsic dimension of the manifold instead of the ambient dimension of the data. Our results demonstrate that deep neural networks can utilize low-dimensional geometric structures of data in machine learning and PDE simulations.

PHILIPPE-ANDRÉ LUNEAU, Université Laval
Conservative Surrogate Models for Optimization with the Active Subspace Method

A way to enforce with high probability nonlinear constraints for optimization using the Active Subspace (AS) method is proposed. The goal of using AS is to lower the dimension of the parametric space of the objective function, reducing effects related to the curse of dimensionality. Generally, this method relies on low-dimensional surrogate models of the objective and the constraints over the AS. Unfortunately, since the surrogate constraints are inexact, this can make the resulting optimal solutions infeasible with respect to the exact constraints. To counter this, an artificial bias is imposed on the training data of the surrogate over the active subspace. Two approaches are proposed to determine the bias: the first one using resampling by bootstrap, and the second one using concentration inequalities. To alleviate the computational cost of bootstrapping, the training data is itself resampled to extract further information about the underlying distribution.

CHRISTOPH ORTNER, UBC
Efficient Parameterization of Many-body Interaction

I will review the atomic cluster expansion (ACE), which provides a systematic, efficient, and interpretable parameterisation of many-body interaction in particle systems. It can be thought of as a method to enlarge the design space of equivariant neural network architectures. ACE is well-suited for parameterising surrogate models of particle systems where it is important to incorporate symmetries and geometric priors into models without sacrificing systematic improvability. The most successful application so far is "learning" interatomic potentials (or, force fields) but the applicability is much broader; it has been adapted to other contexts such as electronic structure (parameterising Hamiltonians), quantum chemistry (wave functions), and elementary particle physics (e.g., jet tagging). The main purpose of my talk will be to explain the framework that enables this breadth of applications, and point out theoretical questions and challenges.

SERGE PRUDHOMME, Polytechnique Montréal
Reduced-order modeling for the wave equation using Green's functions and neural networks

Several deep learning methods have been developed in recent years for the solution of PDE-based problems with the objective of producing techniques that are more flexible and possibly faster than classical discretization approaches. Deep operator networks (DeepONet), for example, aim at solving partial differential equations by learning the inverse of the differential operator for a wide class of input parameters. However, the approach turns out to be expensive for the wave equation at high frequency regimes as the identification of the network parameters may converge slowly. In this talk, we propose an approach based on the representation of the exact solution in terms of the Green's function. The resulting neural network architecture will be referred to as Green operator networks (GreenONets). The novel architecture yields a faster learning and a better generalization error when compared to the classical DeepONet architecture. Performance of the GreenONets and DeepONets will be compared on several numerical examples dealing with wave propagation in homogeneous and heterogeneous media.

ELINA ROBEVA, University of British Columbia
Learning Causal Models via Algebraic Constraints

Abstract: One of the main tasks of causal inference is to learn direct causal relationships among observed random variables. These relationships are usually depicted via a directed graph whose vertices are the variables of interest and whose edges represent direct causal effects. In this talk we will discuss the problem of learning such a directed graph for a linear causal model. We will specifically address the case where the graph may have directed cycles. In general, the causal graph cannot be learned uniquely from observational data. However, in the special case of linear non-Gaussian acyclic causal models, the directed graph can be found uniquely. When cycles are allowed the graph can be learned up to an equivalence class. We characterize the equivalence classes of such cyclic graphs and we propose algorithms for causal discovery. Our methods are based on using specific polynomial relationships which hold among the second and higher order moments of the random vector and which can help identify the graph.

LUANA RUIZ, Johns Hopkins University
Machine Learning on Large-Scale Graphs

Graph neural networks (GNNs) are successful at learning representations from most types of network data but suffer from limitations in large graphs, which do not have the Euclidean structure that time and image signals have in the limit. Yet, large graphs can often be identified as being similar to each other in the sense that they share structural properties. Indeed, graphs can be grouped in families converging to a common graph limit – the graphon. A graphon is a bounded symmetric kernel which can be interpreted as both a random graph model and a limit object of a convergent sequence of graphs. Graphs sampled from a graphon almost surely share structural properties in the limit, which implies that graphons describe families of similar graphs. We can thus expect that processing data supported on graphs associated with the same graphon should yield similar results. In my research, I formalize this intuition by showing that the error made when transferring a GNN across two

graphs in a graphon family is small when the graphs are sufficiently large. This enables large-scale graph machine learning by transference: training GNNs on moderate-scale graphs and executing them on large-scale graphs.

MATTHEW SCOTT, University of British Columbia

When are generative models suitable for signal recovery from subsampled Fourier measurements?

Using the range of generative models as prior sets has shown promise for recovering signals from what appears to be an incomplete set of noisy linear measurements. We present sample complexity bounds when the measurements are subsampled from the rows of a fixed unitary matrix, e.g., subsampled Fourier measurements. To provide meaningful bounds, we introduce a parameter quantifying whether a generative model is well-conditioned with respect to subsampled unitary measurements. We further show how these sample complexity bounds depend on the sampling distribution, and how they can be improved by picking the sampling probabilities in a manner adapted to the generative model.