Mathematics of Machine Learning
Mathématiques de l'apprentissage automatique
(Org: **Ben Adcock** (Simon Fraser University), **Tanya Schmah** (University of Ottawa), **Giang Tran** (University of Waterloo) and/et **Hamid Usefi** (Memorial University))

AARON BERK, McGill University
Variational properties of square root LASSO: Smoothness, uniqueness, explicit solutions

The square root LASSO (SR-LASSO) is a powerful sparse regularization technique widely adopted in statistics, and increasingly popular in the scientific computing and machine learning communities. SR-LASSO is the sum of a data fidelity term and a one-norm weighted by a tuning parameter. It closely resembles the unconstrained formulation of LASSO (Least Absolute Shrinkage and Selection Operator), essentially obtained by "removing" the square from the latter's data fidelity term. This algebraic transformation corresponds with optimal tuning strategies for SR-LASSO that are robust to unknown observation errors. Our goal is to study the (generally, set-valued) solution map of SR-LASSO to determine its sensitivity to the measurement vector and tuning parameter. We present how three increasingly strict assumptions give rise to correspondingly "nice" properties of the SR-LASSO solution map. Our investigation is based on variational analysis, continuing a line of work initiated by the current authors for unconstrained LASSO. First, we show the weakest assumption yields uniqueness for SR-LASSO solutions. The intermediate assumption additionally yields directional differentiability (hence Lipschitzness) of the solution map, as well as an analytic expression for the solution. The final assumption yields continuous differentiability (with respect to the measurement vector and tuning parameter). When the solution is Lipschitz we obtain informative explicit bounds on the Lipschitz constant and contrast this quantification of sensitivity with that for unconstrained LASSO. Compelling numerics flesh out the theoretical discussion.

JASON BRAMBURGER, Concordia University
Auxiliary functions as Koopman observables

Many important statements about dynamical systems can be proved by finding scalar-valued auxiliary functions whose time evolution along trajectories obeys certain pointwise inequalities that imply the desired result. The most familiar of these auxiliary functions is a Lyapunov function to prove steady-state stability, but such functions can also be used to bound averages of ergodic systems, define trapping boundaries, and so much more. In this talk I will highlight a method of identifying auxiliary functions from data using polynomial optimization. The method leverages recent advances in approximating the Koopman operator from data, so-called extended dynamic mode decomposition, to provide system-level information without system identification. The result is a model-agnostic computational method that can be used to bound quantities of interest, develop optimal state-dependent feedback controllers, and discover invariant measures.

KIMON FOUNTOULAKIS, Cheriton School of Computer Science, University of Waterloo
Graph Attention Retrospective

Graph-based learning is a rapidly growing sub-field of machine learning with applications in social networks and bioinformatics. One of the most popular models is graph attention networks. They were introduced to allow a node to aggregate features of neighbour nodes in a non-uniform way, in contrast to simple graph convolution which does not distinguish the neighbours of a node. In this presentation I will discuss multiple results on the performance of graph attention for the problem of node classification for a contextual stochastic block model. The node features are obtained from a mixture of Gaussians and the edges from a stochastic block model. I will show that in an "easy" regime, where the distance between the means of the Gaussians is large enough, graph attention is able to distinguish inter-class from intra-class edges. Thus it maintains the weights of important edges and significantly reduces the weights of unimportant edges. Consequently, I will show that this implies perfect node classification. In the "hard" regime, I will show that every attention mechanism fails to distinguish intra-class from inter-class edges. In addition, I will show that graph attention convolution cannot (almost) perfectly classify the nodes

even if intra-class edges could be separated from inter-class edges. Beyond perfect node classification, I will discuss a positive result on graph attention's robustness against structural noise in the graph. In particular, the robustness result implies that graph attention can be strictly better than both the simple graph convolution and the best linear classifier of node features.

ANASTASIS KRATSIOS, McMaster

A Transfer Principle: Universal Approximators Between Metric Spaces From Euclidean Universal Approximators

We build universal approximators of continuous maps between arbitrary Polish metric spaces X and Y using universal approximators between Euclidean spaces as building blocks. Earlier results assume that the output space Y is a topological vector space. We overcome this limitation by "randomization": our approximators output discrete probability measures over Y. When X and Y are Polish without additional structure, we prove very general qualitative guarantees; when they have suitable combinatorial structure, we prove quantitative guarantees for Hölder-like maps, including maps between finite graphs, solution operators to rough differential equations between certain Carnot groups, and continuous non-linear operators between Banach spaces arising in inverse problems. In particular, we show that the required number of Dirac measures is determined by the combinatorial structure of X and Y. For barycentric Y, including Banach spaces, R-trees, Hadamard manifolds, or Wasserstein spaces on Polish metric spaces, our approximators reduce to Y-valued functions. When the Euclidean approximators are neural networks, our constructions generalize transformer networks, providing a new probabilistic viewpoint of geometric deep learning.

Joint work with: Chong Liu, Matti Lassas, Maarten V. de Hoop, and Ivan Dokmanić

Available at: ArXiV 2304.12231

VINCENT LÉTOURNEAU, University of Ottawa

Complexity measures and regret bounds in reinforcement learning from classical statistical learning theory

I will present recent work that makes use of classical statistical learning theory, specifically complexity measures of supervised learning, to probe the complexity of reinforcement learning problems. We consider a family $\mathcal{M}$ of MDPs over given state and action spaces, and an agent that is sequentially confronted with tasks from $\mathcal{M}$. Although stated for this stepwise change in distributions, the insight we develop is informative for continually changing distributions as well. In order to study how structure of $\mathcal{M}$, viewed as a learning environment, impacts the learning efficiency of the agent, we formulate an RL analog of fat shattering dimension for MDP families and show that this implies a nontrivial lower bound on regret as long as insufficiently many steps have been taken. More precisely, for some constant c which depends on shattering d states, an inexperienced agent that has explored the learning environment for fewer than d steps will necessarily have regret above c on some MDP in the family.

MARTINA NEUMAN, CMSE-MSU

Superiority of GNN over NN in generalizing bandlimited functions

Graph Neural Network (GNN) with its ability to integrate graph information has been widely used for data analyses. However, the expressive power of GNN has only been studied for graph-level tasks but not for node-level tasks, such as node classification, where one tries to interpolate missing nodal labels from the observed ones. In this paper, we study the expressive power of GNN for the said classification task, which is in essence a function interpolation problem. Explicitly, we derive the number of weights and layers needed for a GNN to interpolate a band-limited function in $\mathbb{R}^d$. Our result shows that, the number of weights needed to ϵ-approximate a bandlimited function using the GNN architecture is much fewer than the best known one using a fully connected neural network (NN) - in particular, one only needs $O((\log \epsilon^{-1})^d)$ weights using a GNN trained by $O((\log \epsilon^{-1})^d)$ samples to ϵ-approximate a discretized bandlimited signal in $\mathbb{R}^d$. The result is obtained by drawing a connection between the GNN structure and the classical sampling theorems, making our work the first attempt in this direction.

VAKHTANG PUTKARADZE, University of Alberta

Lie-Poisson Neural Networks

Physics-Informed Neural Networks (PINNs) have acquired a lot of attention in recent years due to their potential for high-performance computations for complex physical systems. The idea of PINNs is to approximate the equations, as well as boundary and initial conditions, through a loss function for a neural network. For applications to canonical Hamiltonian systems, structure-preserving Symplectic Neural Networks (SympNets) were developed, computing canonical transformations and further extended to non-canonical systems due to the application of Darboux's theorem by writing non-canonical systems locally in canonical coordinates. We extend this theory further by developing the Lie-Poisson neural networks (LPNets), which can approximate the motion of solutions on a Poisson manifold given the Poisson bracket. Our method is based on the approximation of the motion using analytically solved motion for test Hamiltonians and given Poisson bracket. The method preserves all Casimirs to machine precision and yields an efficient and promising computational method for the dynamics of several finite-dimensional Lie groups, such as SO(3) (rigid body or satellite), SE(3) (Kirchhoff's equations for underwater vehicle) and other finite-dimensional Lie groups. We also discuss the applications of these ideas to infinite-dimensional systems.

Joint work with Chris Eldred (Sandia National Lab), Francois Gay-Balmaz (CNRS and ENS, France), and Sophia Huraka (U Alberta). The work was partially supported by an NSERC Discovery grant.

TIFFANY VLAAR, McGill University / Mila
Constrained and Multirate Training of Neural Networks

I will describe algorithms for regularizing and training deep neural networks. Soft constraints, which add a penalty term to the loss, are typically used as a form ofexplicit regularization for neural network training. In this talk I describe a method for efficiently incorporating constraints into a stochast ic gradient Langevin framework for the training of deep neural networks. In contrast to soft constraints, our constraints offer direct control of the parameter space, which allows us to study their effect on generalization. In the second part of the talk, I illustrate the presence of latent multiple time scales in deep learning applications. Different features present in the data can be learned by training a neural network on different time scales simultaneously. By choosing appropriate partitionings of the network parameters into fast and slow parts I show that our multirate techniques can be used to train deep neural networks for transfer learning applications in vision and natural language processing in half the time, without reducing the generalization performance of the model.

HAIZHAO YANG, University of Maryland College Park
Finite Expression Method: A Symbolic Approach for Scientific Machine Learning

Machine learning has revolutionized computational science and engineering with impressive breakthroughs, e.g., making the efficient solution of high-dimensional computational tasks feasible and advancing domain knowledge via scientific data mining. This leads to an emerging field called scientific machine learning. In this talk, we introduce a new method for a symbolic approach to solving scientific machine learning problems. This method seeks interpretable learning outcomes in the space of functions with finitely many analytic expressions and, hence, this methodology is named the finite expression method (FEX). It is proved in approximation theory that FEX can avoid the curse of dimensionality in discovering high-dimensional complex systems. As a proof of concept, a deep reinforcement learning method is proposed to implement FEX for learning the solution of high-dimensional PDEs and learning the governing equations of raw data.