Mathematics of machine learning
Mathématiques de l'apprentissage automatique
(Org: **Ben Adcock** (SFU), **Simone Brugiapaglia** (Concordia), **Giang Tran** (Waterloo) and/et **Hamid Usefi** (Memorial))

ANDERSEN ANG, University of Waterloo
Imhomogeous graph signal estimation via a cardinality penalty

A challenge in signal estimation of piecewise smooth signals over a graph is to handle the inhomogeneous levels of smoothness of the signal over the clusters (i.e., communities) of the graph.

We propose a group ℓ_{2-0}-norm-like penalized Graph Trend Filtering (GTF) framework to tackle such inhomogeneity in the graph signal estimation. We prove that solving such penalized GTF is equivalent to jointly performing a k-means clustering on the graph signal (solely based on the signal on nodes, ignoring the graph) and finding a minimum graph cut (solely based on the graph structure, ignoring the signal on nodes), in which the clustering and the cut share the same assignment matrix, indicating that the solution of such GTF graph signal estimation problem is finding a trade-off between k-means clustering on the graph signal and a minimum cut on the graph.

We develop methods (a spectral method and a probabilistic method) to solve such proposed GTF model and present numerical results to support the effectiveness of the methods.

AARON BERK, McGill University
Compressed sensing with generative models and Fourier measurements: provable guarantees under incoherence

In work by Bora et al. (2017), a mathematical framework was developed for compressed sensing guarantees when the measurement matrix is Gaussian and the signal structure is the range of a Lipschitz function (with applications to generative neural networks (GNNs)). We consider measurement matrices derived by sampling uniformly at random rows of a unitary matrix (including subsampled Fourier measurements as a special case). We prove the first known restricted isometry guarantee for compressed sensing with GNNs and subsampled isometries, and provide recovery bounds. Recovery efficacy is characterized by the coherence, a new parameter, which measures the interplay between the range of the network and the measurement matrix. Furthermore, we propose a regularization strategy for training GNNs to have favourable coherence with the measurement operator. We provide compelling numerical simulations that support this regularized training strategy: our strategy yields low coherence networks that require fewer measurements for signal recovery. This, together with our theoretical results, supports coherence as a natural quantity for characterizing generative compressed sensing with subsampled isometries.

QUENTIN BERTRAND, Mila
Synergies Between Disentanglement and Sparsity: a Multi-Task Learning Perspective

Although disentangled representations are often said to be beneficial for downstream tasks, current empirical and theoretical understanding is limited. In this work, we provide evidence that disentangled representations coupled with sparse base-predictors improve generalization. In the context of multi-task learning, we prove a new identifiability result that provides conditions under which maximally sparse base-predictors yield disentangled representations. Motivated by this theoretical result, we propose a practical approach to learn disentangled representations based on a sparsity-promoting bi-level optimization problem. Finally, we explore a meta-learning version of this algorithm based on group Lasso multiclass SVM base-predictors, for which we derive a tractable dual formulation. It obtains competitive results on standard few-shot classification benchmarks, while each task is using only a fraction of the learned representations.

JASON BRAMBURGER, Concordia University
Deep Learning of Conjugate Mappings

Despite many of the most common chaotic dynamical systems being continuous in time, it is through discrete time mappings that much of the understanding of chaos is formed. Henri Poincaré first made this connection by tracking consecutive iterations of the continuous flow with a lower-dimensional, transverse subspace. The mapping that iterates the dynamics through consecutive intersections of the flow with the subspace is now referred to as a Poincaré map, and it is the primary method available for interpreting and classifying chaotic dynamics. Unfortunately, in all but the simplest systems, an explicit form for such a mapping remains outstanding. In this talk I present a method of discovering explicit Poincaré mappings using deep learning to construct an invertible coordinate transformation into a conjugate representation where the dynamics are governed by a relatively simple chaotic mapping. The invertible change of variable is based on an autoencoder, which allows for dimensionality reduction, and has the advantage of classifying chaotic systems using the equivalence relation of topological conjugacies. We illustrate with low-dimensional systems such as the Rössler systems, while also demonstrating the utility of the method on the infinite-dimensional Kuramoto–Sivashinsky equation.

KILIAN FATRAS, Mila - Québec AI Institute, McGill University
Minibatch Optimal Transport distances meets Deep Learning

Optimal transport distances have found many applications in machine learning for their capacity to compare non-parametric probability distributions. Yet their algorithmic complexity generally prevents their direct use on large scale datasets. Among the possible strategies to alleviate this issue, practitioners can rely on computing estimates of these distances over minibatches of data. In this talk, we present an analysis of this practice. We notably argue that it is equivalent to an implicit regularization of the original problem, with appealing properties such as unbiased estimators, gradients and a concentration bound around the expectation. We also highlight in this talk some limits of this strategy, arguing it is not a distance and it can lead to undesirable smoothing effects. As an alternative, we suggest that the same minibatch strategy coupled with unbalanced optimal transport can yield more robust behaviours while preserving the same theoretical properties. Our experimental study shows that in challenging problems associated to domain adaptation, the use of unbalanced optimal transport leads to significantly better results, competing with or surpassing recent baselines.

MARINA GARROTE-LOPEZ, University of British Columbia
Algebraic Optimization of Sequential Decision Problems

In this talk, we study the optimization of the expected long-term reward in finite partially observable Markov decision processes over the set of stationary stochastic policies. We focus on the case of deterministic observations, where the problem is equivalent to optimizing a linear objective subject to quadratic constraints. We characterize the feasible set of this problem as the intersection of a product of affine varieties of rank one matrices and a polytope, which allows us to obtain bounds on the number of critical points of the optimization problem. Finally, we will explain some experiments in which we solve the KKT equations or the Lagrange equations over different boundary components of the feasible set to solve the optimization problem and compare the result to the theoretical bounds and to other constrained optimization methods.

MANUELA GIROTTI, Saint Mary's University
Neural Networks Efficiently Learn Low-Dimensional Representations with SGD

We study the problem of training a two-layer neural network (NN) of arbitrary width using stochastic gradient descent (SGD) where the input $\boldsymbol{x} \in \mathbb{R}^d$ is Gaussian and the target $y \in \mathbb{R}$ follows a multiple-index model, i.e., $y = g(\langle \boldsymbol{u_1}, \boldsymbol{x} \rangle, \ldots, \langle \boldsymbol{u_k}, \boldsymbol{x} \rangle)$ with a noisy link function g. We prove that the first-layer weights of the NN converge to the k-dimensional *principal subspace* spanned by the vectors $\boldsymbol{u_1}, \ldots, \boldsymbol{u_k}$ of the true model, when online SGD with weight decay is used for training. This phenomenon has several important consequences when $k \ll d$. First, by employing uniform convergence on this smaller subspace, we establish a generalization error bound of $O(\sqrt{kd/T})$ after T iterations of SGD, which is independent of the width of the NN. We further demonstrate that, SGD-trained ReLU NNs can learn a single-index target of the form $y = f(\langle \boldsymbol{u}, \boldsymbol{x} \rangle) + \epsilon$ by recovering the principal direction, with a sample complexity linear in d (up to log factors), where f is a monotonic function with at most polynomial growth, and ϵ is the noise. This is in contrast to the known $d^{\Omega(p)}$ sample requirement to learn any degree p polynomial in the kernel regime, and it shows that NNs trained with SGD can outperform the neural tangent kernel at

initialization. Finally, we also provide compressibility guarantees for NNs using the approximate low-rank structure produced by SGD.

This is a joint work with Alireza Mousavi-Hosseini (UofT, Vector), Sejun Park (Korea Univeristy), Ioannis Mitliagkas (UdeM, Mila), and Murat A. Erdogdu (UofT, Vector).

DIANE GUIGNARD, University of Ottawa
Nonlinear approximation of high-dimensional anisotropic analytic functions

The usual approach to model reduction for parametric/random partial differential equations is to construct a linear space of (hopefully small) dimension n which accurately approximates the parameter-to-solution map. This linear reduced model can then be used for various tasks such as building a forward solver or estimating the state or the parameters from data observations. It is well-understood in other problems of numerical computation that nonlinear methods may provide improved numerical efficiency, suggesting the use of nonlinear methods for model reduction as well. In a so-called library approximation, the single linear space is replaced by a collection of affine spaces and the best space may be chosen for each parameter query.

In this talk, we present a specific example of library approximation where the parameter domain is split into a finite number of cells and where different reduced affine spaces of dimension m are assigned to each cell. Given m, we derive an upper bound on the dimension of the library needed to achieve a target accuracy and illustrate the performance of the method through several numerical examples. Finally, we extend this strategy to approximate a general class of anisotropic analytic functions.

RONGJIE LAI, Rensselaer Polytechnic Institute
Learning Manifold-structured Data using Deep networks: Theory and Algorithms

Deep neural networks have made tremendous success in many problems in science and engineering. In this talk, I will discuss our recent efforts on learning non-trivial manifold information hidden in data. Inspired by differential geometry, we propose a Chart Auto-Encoder (CAE) for manifold-structured data representation using a multi-chart latent space. CAE admits desirable manifold properties that auto-encoders with a flat latent space fail to obey. Theoretically, we conduct approximation and nonparametric analysis to understand the proposed CAE. We also verify the effectiveness of the proposed CAE on synthetical and real-world data.

SEBASTIAN MORAGA, Simon Fraser University
Deep neural Networks are effective at learning high-dimensional Banach-valued functions from limited data

Recently, there has been an increasing interest in applying Deep Learning (DL) to computational science and engineering, e.g., computer vision, genetics and computational uncertainty quantification (UQ). In particular, for UQ, high-dimensional problems are often posed in terms of parameterized partial differential equations (PDE) whose solutions take values in abstract spaces. Over the last five years, impressive results have been achieved on such problems using DL techniques, i.e., machine learning based on training Deep Neural Networks (DNN). However, little is known about the efficiency and reliability of DL from the perspectives of stability, robustness, accuracy, and sample complexity. This work focuses on approximating high-dimensional smooth functions taking values in a typically infinite-dimensional Banach space, where training data for such problems is often scarce and may be corrupted by errors. Moreover, obtaining samples is often expensive and involves a complicated black-box PDE solver and high problem dimensionality. Our results provide arguments for DNN approximation of such functions, with both known and unknown parametric dependence, that overcome the main challenge of the curse of dimensionality and account for all sources of error, i.e., sampling, optimization, approximation, and physical discretization. We assert the existence of a class of DNNs with dimension-independent architecture size and training procedures based on minimizing the regularized or unregularized ℓ_2-loss, which achieves near-optimal dimension-independent algebraic convergence rates. We provide numerical results illustrating the practical performance of DNNs on Hilbert-valued functions and preliminary numerical results on Banach-valued functions arising as solutions to parametric PDEs.

CHRISTOPHER MUSCO, New York University
Robust Active Learning via Leverage Score Sampling

Active learning is a promising approach to fitting machine learning models in "data starved" applications, where the cost of collecting data labels is the primary cost of model training. In many of these applications, including in computational science and ML guided engineering, we need active learning methods that work in the challenging agnostic or "adversarial noise" setting. In this setting, collected labels might not match the model being trained, even in expectation. Nevertheless, we seek methods that are robust enough to find the best possible fit with as little data as possible. In this talk, I will discuss recent developments on a flexible class of active learning algorithms based on so-called "leverage score sampling". I will show how leverage score based methods can provably address the challenging agnostic learning problem in a variety of settings, including for linear models, kernel regression models, and also simple neural networks with non-linearities. I will highlight future directions for research and challenging open directions. Based on joint work with Aarshvi Gajjar, Tamás Erdélyi, Chinmay Hegde, Raphael Meyer, Cameron Musco David Woodruff, Taisuke Yasuda, and Samson Zhou.

TAN MINH NGUYEN, University of California, Los Angeles
FourierFormer: Transformer Meets Generalized Fourier Integral Theorem

Multi-head attention empowers the recent success of transformers, the state-of-theart models that have achieved remarkable success in sequence modeling and beyond. These attention mechanisms compute the pairwise dot products between the queries and keys, which results from the use of unnormalized Gaussian kernels with the assumption that the queries follow a mixture of Gaussian distribution. There is no guarantee that this assumption is valid in practice. In response, we first interpret attention in transformers as a nonparametric kernel regression. We then propose the FourierFormer, a new class of transformers in which the dot-product kernels are replaced by the novel generalized Fourier integral kernels. Different from the dot-product kernels, where we need to choose a good covariance matrix to capture the dependency of the features of data, the generalized Fourier integral kernels can automatically capture such dependency and remove the need to tune the covariance matrix. We theoretically prove that our proposed Fourier integral kernels can efficiently approximate any key and query distributions. Compared to the conventional transformers with dot-product attention, FourierFormers attain better accuracy and reduce the redundancy between attention heads. We empirically corroborate the advantages of FourierFormers over the baseline transformers in a variety of practical applications including language modeling and image classification

ESHA SAHA, University of Waterloo
SPADE4: Sparsity and Delay Embedding based Forecasting

Predicting the evolution of diseases is challenging, especially when the data availability is scarce and incomplete. The most popular tools for modelling and predicting infectious disease epidemics are compartmental models. They stratify the population into compartments according to health status and model the dynamics of these compartments using dynamical systems. However, these predefined systems may not capture the true dynamics of the epidemic due to the complexity of the disease transmission and human interactions. In order to overcome this drawback, we propose **Spa**rsity and **D**elay **E**mbedding based **Fore**casting (SPADE4) for predicting epidemics. SPADE4 predicts the future trajectory of an observable variable without the knowledge of the other variables or the underlying system. We use sparsity based random feature model to handle the data scarcity issue and employ Takens' delay embedding theorem to capture the nature of the underlying system from the observed variable. We show that our approach outperforms compartmental models when applied to both simulated and real data.

TANYA SCHMAH, University of Ottawa
Diffeomorphic image matching with a preference for "simple" transformations

Image alignment, i.e. registration, is a fundamental problem in computer vision, including in medical imaging, where it allows comparison of images from different subjects or different times. While deep learning has made an important impact on this problem, the gold standard is still the geometric, or variational, approach which is based on geodesic flows in a diffeomorphism

group (or in the group orbit of a particular image). A right-invariant Riemannian metric is used both to define the geodesics and to regularize the optimization problem by penalizing larger deformations.

We consider variants of diffeomorphic image registration that prefer "simple" deformations, defined as those in a pre-specified subgroup G, for example the affine group or a projective linear group. One approach is to use a Riemannian metric (or degenerate metric) that penalizes velocities tangent to G only very mildly (or not at all). While theoretically satisfying, this makes computing geodesics more difficult, so we also consider flows of fixed vector fields.

CHRYSTAL SMITH, York University

Natural Language Processing in the field of Medical Translation

Advances in artificial intelligence and machine learning, such as deep learning neural networks, embed syntactic and semantic information using vectors to achieve accuracy of response and human understanding of discourse and sentiment for Natural Language Processing (NLP). Despite its use in the field of medical translation the mathematical foundations of this approach is not well understood and lingering problems persist. In this talk I review current methods and consider how Statistical NLP systems and neural networks produce natural language for the field of medical translation.

WEIQI WANG, Concordia University

Compressive Fourier collocation methods for high-dimensional diffusion equations with periodic boundary conditions.

High-dimensional Partial Differential Equations (PDEs) are a popular mathematical modelling tool. However, standard numerical techniques for solving High-dimensional PDEs are typically affected by the curse of dimensionality. In this work, we tackle this challenge while focusing on stationary diffusion equations defined over a high-dimensional domain with periodic boundary conditions. Inspired by recent progress in high-dimensional sparse function approximation, we propose a new method called compressive Fourier collocation. Combining ideas from compressive sensing and spectral collocation, our method uses Monte Carlo sampling and employs sparse recovery techniques, such as orthogonal matching pursuit and l^1 minimization, to approximate the Fourier coefficients on given index sets of the PDE solution. We conduct a rigorous theoretical analysis showing that the approximation error of the proposed method is comparable with the best s-term approximation (with respect to the Fourier basis) to the solution and mitigates the curse of dimensionality with respect to the number of collocation points under sufficient conditions on the regularity of the diffusion coefficient. We present numerical experiments that illustrate the accuracy and stability of the method for the approximation of sparse and compressible solutions. In our current work, noticing that a bottleneck towards improving the solution accuracy is the choice of the index set, we develop a method using orthogonal matching pursuit to adaptively select the elements of the index set. In addition, we seek an efficient neural network model to solve the high-dimensional PDE, with the goal of comparing the performance of the adaptive method with a deep learning-based approach.