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

AVI GUPTA, Simon Fraser University

MOHAMED HIBAT-ALLAH, University of Waterloo

SPENCER HILL, Queen's University

SHIKHAR JAISWAL, University of Toronto

ANASTASIS KRATSIOS, McMaster University

SOPHIE MORIN, Polytechnique Montreal

RACHEL MORRIS, Concordia University

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

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

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.