## Mathematical Foundations of Machine Learning
## Fondations mathématiques de l'apprentissage automatique

(Org: **Ben Adcock** (Simon Fraser University) and/et **Simone Brugiapaglia** (Concordia))

**SHAI BEN-DAVID**, University of Waterloo

*Models of statistical learning and their learnability characterizing dimensions*

I will define a model for general statistical learning that captures as specific cases most of the common machine learning tasks, and discuss some natural special cases of that model. The talk will focus on the existence of notions of dimensions that characterize PAC-style learnability in these models. I will propose a couple of possible definitions of such notions of dimensions. For some, such as binary classification, there are well known learnability characterizing dimensions (e.g., the VC-dimension). However, for the most general model we have shown that assuming ZFC is consistent, no such dimension exists. For some other interesting variants, the existence of learnability-characterization dimension is an open question.

**AARON BERK**, Concordia University

*Sensitivity to parameter selection for LASSO programs*

Compressed sensing theory explains why LASSO programs recover structured high-dimensional signals with minimax order-optimal error. Yet, the optimal choice of the program's governing parameter is often unknown in practice. It is still unclear how variation of the governing parameter impacts recovery error in compressed sensing, which is otherwise provably stable and robust. We provide an overview of parameter sensitivity in LASSO programs in the setting of proximal denoising; and of compressed sensing with subgaussian measurement matrices and gaussian noise. We demonstrate how two popular ell-1-based minimization programs exhibit sensitivity with respect to their parameter choice and illustrate the theory with numerical simulations. For example, a 1% percent error in the estimate of a parameter can cause the error to increase by a factor of $10^9$, while choosing a different LASSO program avoids such sensitivity issues. We hope that revealing parameter sensitivity regimes of LASSO programs helps to inform a practitioner's choice.

**ALEX BIHLO**, Memorial University of Newfoundland

*A multi-model physics-informed neural network approach for solving the shallow-water equations on the sphere*

Multi-model physics-informed neural networks are developed for solving the shallow-water equations on the sphere. Vanilla physics-informed neural networks are trained to satisfy differential equations along with the prescribed initial and boundary data, and thus can be seen as an alternative approach to solving differential equations compared to traditional numerical approaches such as finite difference, finite volume or spectral methods. I will discuss the training difficulties of vanilla physics-informed neural networks for the shallow-water equations on the sphere and propose a simple multi-model approach to tackle test cases of comparatively long time intervals. I will illustrate the abilities of the method by solving the most prominent test cases proposed by Williamson et al. [J. Comput. Phys. 102, 211-224, 1992]. This is joint work with Roman O. Popovych.

**CLAIRE BOYER**, LPSM, Sorbonne Université Paris

*Some ways of missing data handling in machine learning*

One of the big ironies of data sciences is that the more data we have, the more missing data are likely to appear. After discussing the various issues presented by the missing data in daily-life machine learning, we will present different ways to tackle them for different purposes: (i) one may want to infer the missing values, this is what is called imputation. Imputation can be performed with low-rank techniques, but with optimal transport as well; (ii) one may want to handle missing values in regression, both for performing model estimation or for predictive concern; (iii) one may want to cope with missing values in a non-supervised learning scenario such as for clustering data. We will present some insights and works trying to address the previous issues.

**NICK DEXTER**, Simon Fraser University
*Improving efficiency of deep learning approaches for scientific machine learning*

Sparse reconstruction techniques from compressed sensing have been successfully applied to many application areas, including signal processing, inverse problems in imaging, and approximation of solutions to parameterized partial differential equations (PDE). Such approaches are capable of exploiting the sparsity of the signal to achieve highly accurate approximations with minimal sample complexity. For problems whose solutions possess a great deal of structure, their recovery properties can be further enhanced through a combination of carefully selected weighting or structured sampling schemes. Recently connections between compressed sensing and deep learning have been explored, and the existence of deep neural network (DNN) architectures which achieve the same sample complexity and accuracy as compressed sensing on function approximation problems has been established. In this work, we further explore these connections and sparse neural network approximation in the context of high-dimensional parameterized PDE problems. We provide a full error analysis for such problems, explicitly accounting for the errors of best approximation (describing DNN expressibility), spatial discretization of the PDE, and the algorithm used in solving the underlying optimization problem. We complement our theoretical contributions with detailed numerical experiments, demonstrating the potential for sparse neural network approximation in scientific machine learning contexts.

**AXEL FLINTH**, Chalmers University of Technology
*A universal rotation equivariant and permutation invariant neural network architecture*

A function is equivariant to a group action if a transformation of the input results in a similar transformation of the input. In this talk, we consider the action of the rotation group on 2D point clouds. Since permutation of the points in a cloud leaves it invariant, this means that we are dealing with functions that are permutation invariant and rotation equivariant. In this talk, we describe a simple neural network architecture which is capable of universally approximating such functions. The talk is based on joint work with Georg Bökman and Fredrik Kahl.

**BAMDAD HOSSEINI**, University of Washington
*Solving nonlinear PDEs with Gaussian Processes*

I present a simple, rigorous, and interpretable framework for solution of nonlinear PDEs based on the framework of Gaussian Processes. The proposed approach provides a natural generalization of kernel methods to nonlinear PDEs; has guaranteed convergence; and inherits the state-of-the-art computational complexity of linear solvers for dense kernel matrices. I will outline our approach by focusing on an example nonlinear elliptic PDE followed by further numerical examples and discussion of some theory.

**AUKOSH JAGANNATH**, University of Waterloo
*Online SGD on non-convex losses from high-dimensional inference*

Stochastic gradient descent (SGD) is a popular tool in data science. Here one produces an estimator of an unknown parameter from independent samples of data by iteratively optimizing a loss function, which is random and often non-convex. We study the performance of SGD from an uninformative (random) start in the setting where the parameter space is high-dimensional. We develop nearly sharp thresholds for the number of samples needed for consistent estimation as one varies the dimension. They depend only on an intrinsic property of the population loss, called the information exponent and do not assume uniform control on the loss itself (e.g., convexity or Lipschitz-type bounds). These thresholds are polynomial in the dimension and the precise exponent depends explicitly on the information exponent. As a consequence, we find that except for the simplest tasks, almost all of the data is used in the initial search phase, i.e., just to get non-trivial correlation with the ground truth, and that after this phase, the descent is rapid and exhibits a law of large numbers. We illustrate our approach by applying it to a wide set of inference tasks such as parameter estimation for generalized linear models and spiked tensor models, phase retriveal, online PCA, as well as supervised learning for single-layer networks with general activation functions. Joint work with G. Ben Arous (NYU Courant) and R. Gheissari (Berkeley)

**COURTNEY PAQUETTE**, McGill University
*SGD in the Large: Average-case Analysis, Asymptotics, and Stepsize Criticality*

In this talk, I will present a framework, inspired by random matrix theory, for analyzing the dynamics of stochastic gradient descent (SGD) when both the number of samples and dimensions are large. Using this new framework, we show that the dynamics of SGD on a least squares problem with random data become deterministic in the large sample and dimensional limit. Furthermore, the limiting dynamics are governed by a Volterra integral equation. This model predicts that SGD undergoes a phase transition at an explicitly given critical stepsize that ultimately affects its convergence rate, which we also verify experimentally. Finally, when input data is isotropic, we provide explicit expressions for the dynamics and average-case convergence rates. These rates show significant improvement over the worst-case complexities.

**LAURA THESING**, LMU Munich
*Which networks can be learned by an algorithm? – Expressivity meets Turing in Deep Learning*

Deep learning with neural networks celebrates a big success in a wide range of applications, with performance exceeding what many thought would be possible. However, we are also aware of the inherent instabilities to small perturbations of the input. Methods to show these instabilities with adversarial attacks were first introduced for image classification tasks and were later extended to malware detection, sound detection, text analysis, medical fraud, and inverse problems. The instabilities might be surprising when considering the results about the expressivity of neural networks that show that the class is rich enough for continuous network approximations. Already the universal approximation theorem shows that all continuous functions can be approximated with neural networks. Hence given a continuous problem it can be approximated by stable neural networks. Therefore, the question arises: Why are the learned network approximations typically unstable when stable neural networks exist? One aspect that is often overlooked is the computability of neural networks. Indeed, there are examples where stable neural networks exist, but no algorithm can compute them. This fact leads to the questions about which networks are computable functions and for which networks do we have an algorithm that can map from point samples to the approximation. We will elaborate on these questions in this talk.

**GIANG TRAN**, University of Waterloo
*Adaptive Group Lasso for Time-Dependent Data*

In this paper, we propose an adaptive group Lasso deep neural network for high-dimensional function approximation where input data are generated from a dynamical system and the target function depends on a few active variables or a few linear combinations of variables. We approximate the target function by a deep neural network and enforce an adaptive group Lasso constraint to the weights of a suitable hidden layer in order to represent the constraint on the target function. We prove the total loss decay and study the convergence analysis of the proposed framework. Our empirical studies show that the proposed method outperforms recent state-of-the-art methods including the sparse dictionary matrix method, neural networks with or without group Lasso penalty.

**SOLEDAD VILLAR**, Johns Hopkins University
*Equivariant machine learning structured like classical physics*

There has been enormous progress in the last few years in designing neural networks that respect the fundamental symmetries and coordinate freedoms of physical law. Some of these frameworks make use of irreducible representations, some make use of high-order tensor objects, and some apply symmetry-enforcing constraints. Different physical laws obey different combinations of fundamental symmetries, but a large fraction (possibly all) of classical physics is equivariant to translation, rotation, reflection (parity), boost (relativity), and permutations. Here we show that it is simple to parameterize universally approximating polynomial functions that are equivariant under these symmetries, or under the Euclidean, Lorentz, and Poincaré groups, at any dimensionality d. The key observation is that nonlinear O(d)-equivariant (and related-group-equivariant) functions can be universally expressed in terms of a lightweight collection of scalars – scalar products and scalar contractions of the scalar, vector, and tensor inputs. We complement our theory with numerical examples that show that the scalar-based method is simple, efficient, and scalable.