INVERSE PROBLEMS AND MACHINE LEARNING
The California Institute of Technology
February 9 – 11, 2018

Abstracts

Jens Behrmann
University of Bremen

TOWARDS UNDERSTANDING THE ILL-POSEDNESS OF INVERTING RECTIFIER NETWORKS

First, we motivate the inversion of deep networks by discussing different scenarios and approaches. In order to better understand potential issues arising during inversion, we analyze pre-images of relu-layers and the stability of the inverse of relu-networks. In particular, this leads to conditions when pre-images of a relu-layer are a point, finite or infinite. Furthermore, we simplify the stability analysis by leveraging the locally-linear nature of relu-networks. Combining theoretical and numerical results, we show which mechanism affect the inverse. In the end, we present applications to input reconstruction for convolutional networks.

———— ◇ ————

Misha Belkin
Ohio State University

MAKING SHALLOW LEARNING GREAT AGAIN

———— ◇ ————

Joan Bruna
New York University

ON THE LOSS SURFACE OF NEURAL NETWORKS

The loss surface of deep neural networks has recently attracted interest in the optimization and machine learning communities as a prime example of high-dimensional non-convex problem. In this talk, we study conditions on the data distribution and model architecture that prevent the existence of bad local minima. We first take a topological approach and characterize absence of bad local minima by studying the connectedness of the loss surface level sets. Our work quantifies and formalizes the interplay between the smoothness of the data distribution and model over-parametrization. Our main theoretical contribution is to prove that half-rectified single layer networks are asymptotically connected, and we provide explicit bounds that reveal the aforementioned interplay.

The conditioning of gradient descent is the next challenge we address. We study this question through the geometry of the level sets, and we introduce an algorithm to efficiently estimate the regularity of such sets on large-scale networks. Our empirical results show that these level sets remain connected throughout all the learning phase, suggesting a near convex behavior, but they become exponentially more curvy as the energy level decays, in accordance to what is observed in practice with very low curvature attractors.

Joint work with Daniel Freeman (UC Berkeley), Luca Venturi and Afonso Bandeira (Courant, NYU).

———— ◇ ————

Cristoph Brune
University of Twente

DEEP LEARNING THEORY WITH APPLICATION TO CANCER RESEARCH

———— ◇ ————

Venkat Chandrasekaran
California Institute of Technology

LEARNING REGULARIZERS FROM DATA

Regularization techniques are widely employed in the solution of inverse problems in data analysis and scientific computing due to their effectiveness in addressing difficulties due to ill-posedness. In their most common manifestation, these methods take the form of penalty functions added to the objective in optimization-based approaches for solving inverse problems. The purpose of the penalty function is to induce a desired structure in the solution, and these functions are specified based on prior domain-specific expertise. We consider the problem of learning suitable regularization functions from data in settings in which precise domain knowledge is not directly available; the objective is to identify a regularizer to promote the type of structure contained in the data. The regularizers obtained using our framework are specified as convex functions that can be computed efficiently via semidefinite programming. Our approach for learning such semidefinite regularizers combines recent techniques for rank minimization problems along with the Operator Sinkhorn iteration. (Joint work with Yong Sheng Soh)

———— ◇ ————

Pratik Chaudhari
UCLA

UNRAVELING THE MYSTERIES OF STOCHASTIC GRADIENT DESCENT ON DEEP NETWORKS

Stochastic gradient descent (SGD) is widely believed to perform implicit regularization when used to train deep neural networks. The precise manner in which this occurs has thus far been elusive. I will show that SGD solves a variational optimization problem: it minimizes an average potential over the posterior distribution of weights along with an entropic regularization term.

I will show that due to highly non-isotropic mini-batch gradient noise for deep networks, the above potential is not the original loss function; SGD minimizes a different loss than the one it computes its gradients on. More surprisingly, SGD does not even converge in the classical sense: most likely trajectories of SGD for deep networks do not behave like Brownian motion around critical points. Instead, they are limit cycles with deterministic dynamics in the weight space, far away from critical points of the original loss. I will also discuss connections of such “out-of-equilibrium” behavior of SGD with the generalization performance of deep networks.

———— ◇ ————

Matthew Dunlop
Caltech

UQ IN GRAPH-BASED CLASSIFICATION
(joint with Luo)

———— ◇ ————

Nicolas Flammarion
UC, Berkeley

OPTIMAL RATES FOR LEAST-SQUARES REGRESSION THROUGH SGD

———— ◇ ————

Rémi Gribonval
Inria

LEARNING FROM RANDOM MOMENTS

The talk will outline the main features of a recent framework for large-scale learning called compressive statistical learning. Inspired by compressive sensing, the framework allows drastic volume and dimension reduction to learn from large/distributed/streamed data collections . Its principle is to compute a low-dimensional (nonlinear) sketch (a vector of random empirical generalized moments), in essentially one pass on the training collection. For certain learning problems, small sketches have been shown to capture the information relevant to the considered learning task, and empirical learning algorithms have been proposed to learn from such sketches. As a proof of concept, more than a thousands hours of speech recordings can be distilled to a sketch of only a few kilo-bytes, while capturing enough information estimate a Gaussian Mixture Model for speaker verification. The framework, which is endowed with statistical guarantees in terms of learning error, will be illustrated on sketched clustering, and sketched PCA, using empirical algorithms inspired by sparse recovery algorithms used in compressive sensing. Finally, we will discuss the promises of the framework in terms of privacy-aware learning, and its connections with information preservation along pooling layers of certain convolutional neural networks.

Joint work with Nicolas Keriven (ENS Paris, France), Yann Traonmilin (Univ Bordeaux, France) and Gilles Blanchard (Universität Potsdam, Germany).

———— ◇ ————

Eldad Haber
University of British Columbia

DEEP NEURAL NETWORKS MEET PARTIAL DIFFERENTIAL EQUATIONS

In this talk, we will explore deep neural networks from a dynamical systems point of view. We will show that the learning problem can be cast as a path planning problem with PDE constraint. This opens the door to conventional Computational techniques that can speed up the learning process and avoid some of the local minima.

———— ◇ ————

Jiequn Han
Princeton University

SOLVING HIGH-DIMENSIONAL PARTIAL DIFFERENTIAL EQUATIONS USING DEEP LEARNING

Developing algorithms for solving high-dimensional partial differential equations (PDEs) has been an exceedingly difficult task for a long time, due to the notoriously difficult problem known as the ``curse of dimensionality''. In this talk, we introduce a new approach based on deep learning, deep BSDE method, to solve general high-dimensional parabolic PDEs. To this end, the PDEs are reformulated as a control theory problem and the gradient of the unknown solution is approximated by neural networks, very much in the spirit of deep reinforcement learning with the gradient acting as the policy function. Numerical results of a variety of examples demonstrate that the proposed algorithm is quite effective in high-dimensions, in terms of both accuracy and speed.

———— ◇ ————

Nikola Kovachki
Caltech

DERIVATIVE-FREE ENSEMBLE METHODS FOR MACHINE LEARNING TASKS

The standard probabilistic perspective of machine learning devises tasks as empirical risk-minimizations problems that are usually solved by stochastic gradient decent (SGD). We present an alternate formulation of these tasks as classical inverse or filtering problems. In particular, we focus on offline and online supervised learning with deep neural networks as well as graph-based semi-supervised learning. Furthermore we propose an efficient, gradient-free algorithm for finding a solution to these problems based on the classical ensemble Kalman filter (EnKF). The essence of the procedure is discretizing the continuous time limit of EnKF when viewed as an approximation to a sequential Monte Carlo method. We suggest several modifications to the original limit based on recent analysis of the linear case as well as empirically successful heuristics of SGD. Numerical results demonstrate a wide applicability and robustness of the proposed algorithm.

———— ◇ ————

Dirk Lorenz
TU Braunschweig

RANDOMIZED SPARSE KACZMARZ METHODS

Randomized incremental methods are important in several learning problems as they allow to treat small batches of possibly very large objective functions individually. A classical method of this type is the Kaczmarz method for linear systems. We will introduce a sparse version of this method that allows to either calculate sparse solutions of underdetermined systems or approximate sparse solutions of overdetermined system more quickly than the usual Kaczmarz method. Moreover we will show that the randomized version of the method converges linearly.

———— ◇ ————

Xiyang Luo
UCLA

UQ IN GRAPH-BASED CLASSIFICATION
(joint with Dunlop)

———— ◇ ————

Mauro Maggioni
Johns Hopkins University

LEARNING EFFECTIVE DIFFUSION PROCESSES ON MANIFOLDS

We discuss a geometry-based statistical learning framework for performing model reduction and modeling of stochastic high-dimensional dynamical systems. We consider two complementary settings. In the first one, we are given long trajectories of a system, e.g. from molecular dynamics, and we discuss new techniques for estimating, in a robust fashion, an effective number of degrees of freedom of the system, which may vary in the state space of then system, and a local scale where the dynamics is well-approximated by a reduced dynamics with a small number of degrees of freedom. We then use these ideas to produce an approximation to the generator of the system and obtain, via eigenfunctions of an empirical Fokker-Planck question, reaction coordinates for the system that capture the large time behavior of the dynamics. We present various examples from molecular dynamics illustrating these ideas. In the second setting we only have access to a (large number of expensive) simulators that can return short simulations of high-dimensional stochastic system, and introduce a novel statistical learning framework for learning automatically a family of local approximations to the system, that can be (automatically) pieced together to form a fast global reduced model for the system, called ATLAS. ATLAS is guaranteed to be accurate (in the sense of producing stochastic paths whose distribution is close to that of paths generated by the original system) not only at small time scales, but also at large time scales, under suitable assumptions on the dynamics. We discuss applications to homogenization of rough diffusions in low and high dimensions, as well as relatively simple systems with separations of time scales, and deterministic chaotic systems in high-dimensions, that are well-approximated by stochastic differential equations.

———— ◇ ————

Michael Mahoney
UC, Berkeley

SECOND ORDER MACHINE LEARNING

A major challenge for large-scale machine learning, and one that will only increase in importance as we develop models that are more and more domain-informed, involves going beyond high-variance first-order optimization methods to more robust second order methods. Here, we consider the problem of minimizing the sum of a large number of functions over a convex constraint set, a problem that arises in many data analysis, machine learning, and more traditional scientific computing applications, as well as non-convex variants of these basic methods. While this is of interest in many situations, it has received attention recently due to challenges associated with training so-called deep neural networks. We establish improved bounds for algorithms that incorporate sub-sampling as a way to improve computational efficiency, while maintaining the original convergence properties of these algorithms. These methods exploit recent results from Randomized Linear Algebra on approximate matrix multiplication. Within the context of second order optimization methods, they provide quantitative convergence results for variants of Newton's methods, where the Hessian and/or the gradient is uniformly or non-uniformly sub-sampled, under much weaker assumptions than prior work. We also discuss extensions of the basic method to trust region and cubic regularization algorithms for non-convex optimization problems, interesting empirical observations on both convex and non-convex problems, as well as several non-obvious extensions.

———— ◇ ————

Hrushikesh Mhaskar
Claremont Graduate University

MACHINE LEARNING MEETS SUPER-RESOLUTION

We demonstrate a duality between the problems of function approximation in machine learning and super-resolution. Using the sphere as a case study, we demonstrate how the problems related to the degree of approximation can be interpreted as problems of accuracy in the recovery of measures supported on continua, and the same tools can be used to address both these sets of problems.

———— ◇ ————

Adam Oberman
McGill University

CONTINUOUS TIME METHODS FOR LARGE SCALE OPTIMIZATION

Optimization in large scale machine learning problems is restricted to first order methods, such as gradient descent or stochastic gradient descent. Acceleration methods, such as Nesterov's, converge faster, using the same gradient oracle. In this talk, we explain how ideas from continuous time dynamics can be used to understand and generate accelerated optimization algorithms. We will make connections between ODE methods, e.g. explicit and implicit method, and proximal point methods, and show how, in the case of stochastic gradients, this corresponds to the Local Entropy algorithm. We will also show how to use Liapunov function theory for ODEs to obtain explicit methods for gradient descent which give the corresponding rate.

———— ◇ ————

Stanley Osher
UCLA

PDE BASED APPROACHES TO NONCONVEX OPTIMIZATION

We will draw from some of our new results in optimization, often related to partial differential equations, to improve performance of algorithms ranging from data dependent activation in deep learning, training quantized neural networks, optimizing neural networks solving the phase lift problem and diagnosing forward operator error. (Joint with many people.)

———— ◇ ————

Braxton Osting
University of Utah

A GENERALIZED MBO DIFFUSION GENERATED METHOD FOR CONSTRAINED HARMONIC MAPS

A variety of tasks in inverse problems and data analysis can be formulated as the variational problem of minimizing the Dirichlet energy of a function that takes values in a certain submanifold and possibly satisfies additional constraints. These additional constraints may be used to enforce fidelity to data or other structural constraints arising in the particular problem considered. I'll present a generalization of the Merriman-Bence-Osher (MBO) method for minimizing such a functional. I'll give examples of how this method can be used for the geometry processing task of generating quadrilateral meshes, finding Dirichlet partitions, and constructing smooth orthogonal matrix valued functions. For this last problem, I'll prove the stability of the method by introducing an appropriate Lyapunov function, generalizing a result of Esedoglu and Otto to matrix-valued functions. I'll also state a convergence result for the method. This is joint work with Dong Wang and Ryan Viertel.

———— ◇ ————

Sergiy Pereverzyev, Jr.
University of Innsbruck

REGULARIZED INTEGRAL OPERATORS IN TWO-SAMPLE PROBLEM

In the two-sample problem, which is a problem of data mining, one has two samples of observations drawn from two unknown probability measures, and one uses them to decide whether these two measures are equal or not. At the same time, for any probability measures, one can also consider the corresponding integral covariance operators on a Reproducing Kernel Hilbert Space (RKHS). Such operators seem to be powerful tools for investigating the two-sample problem. For example, as we observe, a recently proposed Maximum Mean Discrepancy (MMD) approach to this problem is in fact based on the evaluation of covariance operators at only one particular element. This observation suggests a more deeper exploration of the covariance operators for solving the two-sample problem. In this talk, we are going to demonstrate how these operators can be used to formulate several effective two-sample tests. We will illustrate the proposed tests with numerical examples.

This work is supported by the Austrian Science Fund (FWF): project P 29514-N32.

———— ◇ ————

Ekaterina Rapinchuk
Michigan State University

AN AUCTION APPROACH TO SEMI-SUPERVISED DATA CLASSIFICATION

We reinterpret the semi-supervised data classification problem using an auction dynamics framework (inspired by real life auctions) in which elements of the data set make bids to the class of their choice. This leads to a novel forward and reverse auction method for data classification that readily incorporates volume/class-size constraints into an accurate and efficient algorithm requiring remarkably little training/labeled data. We prove that the algorithm is unconditionally stable, and state its average and worst case time complexity.

———— ◇ ————

Lorenzo Rosasco
Massachusetts Institute of Technology

AN INVERSE PROBLEM PERSPECTIVE ON MACHINE LEARNING

———— ◇ ————

Johannes Schmidt-Hieber
Leiden University

STATISTICAL THEORY FOR DEEP NEURAL NETWORKS WITH RELU ACTIVATION FUNCTION

The universal approximation theorem states that neural networks are capable of approximating any continuous function up to a small error that depends on the size of the network. The expressive power of a network does, however, not guarantee that deep networks perform well on data. For that, control of the statistical estimation risk is needed. In the talk, we derive statistical theory for fitting deep neural networks to data generated from the multivariate nonparametric regression model. It is shown that estimators based on sparsely connected deep neural networks with ReLU activation function and properly chosen network architecture achieve the minimax rates of convergence (up to logarithmic factors) under a general composition assumption on the regression function. The framework includes many well-studied structural constraints such as (generalized) additive models. While there is a lot of flexibility in the network architecture, the tuning parameter is the sparsity of the network. Specifically, we consider large networks with number of potential parameters being much bigger than the sample size. Interestingly, the depth (number of layers) of the neural network architectures plays an important role and our theory suggests that scaling the network depth with the logarithm of the sample size is natural.

———— ◇ ————

Dejan Slepcev
Carnegie Mellon University

REGULARIZING OBJECTIVE FUNCTIONALS OF SEMI-SUPERVISED LEARNING

———— ◇ ————

Stefano Soatto
UCLA/AWS

THE EMERGENCE THEORY OF DEEP LEARNING: PERCEPTION, INFORMATION THEORY AND PAC BAYES

Theories of Deep Learning are like anatomical parts best not named explicitly: Everyone seems to have one. That is why it is important for a theory to be inclusive: It has to be compatible with all known empirical results, and at the very least explain observed empirical phenomena, if not predict new ones. I will describe the basic elements of the Emergence Theory of Deep Learning, that started as a general theory for representations, and is comprised of three parts: (1) Formalization of desirable properties a representation should possess, based on classical principles of statistical decision and information theory: Sufficiency, Invariance, Minimality, Independence. This has nothing to do with Deep Learning, but is closely tied with the notion of Information Bottleneck and Variational Inference. (2) Description of common empirical losses employed in Deep Learning (e.g., empirical cross-entropy), and implicit or explicit regularization practices, including Dropout, Pooling, as well as recently discovered additive entropic components of the loss computed by stochastic gradient descent (SGD). Finally, (3) theorems and bounds that show that minimizing suitably (implicitly or explicitly) regularized losses with SGD with respect of the weights implies optimization of the loss described in (1) with respect to the activations of a deep network, and therefore achievement of the desirable properties of the resulting representation formalized in (1). The link between the two is specific to the architecture of deep networks. The theory is related to the Information Bottleneck, but not that described in recent theories, but instead a new Information Bottleneck for the weights of a network, rather than the activation. It is also related to PAC-Bayes, and could be derived with that lens, providing independent validation. It is also related to Kolmogorov complexity.

———— ◇ ————

Bharath Sriperumbudur
Pennsylvania State University

ON APPROXIMATE KERNEL PCA USING RANDOM FEATURES

Kernel methods are powerful learning methodologies that provide a simple way to construct nonlinear algorithms from linear ones. Despite their popularity, they suffer from poor scalability in big data scenarios. Various approximation methods, including random feature approximation, have been proposed to alleviate the problem. However, the statistical consistency of most of these approximate kernel methods is not well understood except for kernel ridge regression wherein it has been shown that the random feature approximation is not only computationally efficient but also statistically consistent with a minimax optimal rate of convergence. In this work, we investigate the efficacy of random feature approximation in the context of kernel principal component analysis (KPCA) by studying the statistical behavior of approximate KPCA in terms of the convergence of eigenspaces and the reconstruction error.

———— ◇ ————

Pengchuan Zhang
Microsoft Research AI

ANALYSIS AND APPLICATIONS OF DEEP GENERATIVE MODELS

Deep generative models trained by generative adversarial networks (GANs) have achieved impressive success in approximating complex high-dimensional distributions, like the distribution of natural images. By viewing GANs as minimizing certain moment matching loss over a set of discriminators, we study the discrimination-generation tradeoff when designing the discriminator set: the discriminator set should be large enough to be able to uniquely identify the true distribution (discriminative), and be small enough to go beyond memorizing samples (generalizable). We show that a discriminator set is guaranteed to be discriminative whenever its linear span is dense in the set of bounded continuous functions. This is a very mild condition satisfied even by neural networks with a single neuron. Further, we develop generalization bounds between the learned distribution and target distribution under different evaluation metrics. We apply GANs to the task of text-to-image synthesis and show that the learned deep generative models are able to generate realistic and diverse images. In the setting of Bayesian inverse problems, we utilize deep generative networks to approximate the posterior distributions, and design new training algorithms (different from GANs) to train the deep generative networks.