Workshop on Mathematics of Deep Learning

Weierstrass Institute
for Applied Analysis and Stochastics

December 3 – 5, 2019

www.wias-berlin.de/workshops/DL2019/

Organizers: Martin Eigel (WIAS Berlin)
Peter Friz (TU Berlin/WIAS Berlin)
Reinhold Schneider (TU Berlin)
Volodia Spokoiny (HU Berlin/WIAS Berlin)

Weierstraß-Institut für
Angewandte Analysis und Stochastik

www.wias-berlin.de

supported by MATH+

Tuesday, December 3, 2019	
13:00 - 13:40	REGISTRATION & COFFEE
13:40	OPENING
13:45	**Ingo Steinwart (Stuttgart)** Some thoughts and questions towards a statistical understanding of DNNs
14:30	**Dmitry Yarotsky (Moscow)** The fast non-classical approximation phase in deep neural networks
15:15 - 15:45	COFFEE BREAK
15:45	**Philipp Petersen (Vienna)** High-dimensional approximation by deep neural networks in the context of parametric PDEs
16:30	**Andrea Manzoni (Milan)** Nonlinear dimensionality reduction of parametrized PDEs using deep learning techniques
16:50	**Ilja Klebanov (Berlin)** Rigorous theory of conditional mean embeddings

Wednesday, December 4, 2019	
09:00	**Christoph Schwab (Zurich)** Deep neural network expression for PDEs with uncertain input and data
09:45	**Helmut Bölcskei (Zurich)** Fundamental limits of deep neural network learning
10:30 – 11:00	COFFEE BREAK
11:00	**Gitta Kutyniok (Berlin)** Transferability of spectral graph convolutional neural networks
11:45	**Ivan Oseledets (Moscow)** Greedy algorithm for learning deep neural network architectures
12:30 – 14:00	LUNCH BREAK
14:00	**Edwin Stoudenmire (New York)** Theory of a generative machine learning algorithm based on matrix product states
14:45	**Johannes Rauh (Leipzig)** Approximation properties of stochastic neural networks
15:05 – 15:30	COFFEE BREAK
15:30	**Patrick Gelß (Berlin)** Tensor-based algorithms for image classification
15:50	**Philipp Trunschke (Berlin)** Image classification with tensor networks
16:10	**Leon Sallandt (Berlin)** A near model-free method for solving the Hamilton-Jacobi-Bellman equation in high dimensions
19:00 – 21:00	DINNER at “Umspannwerk OST”, Palisadenstr. 48, 10243 Berlin

Thursday, December 5, 2019	
09:00	**Arnulf Jentzen (Münster)** High-dimensional approximation capacities of deep neural networks
09:45	**Philipp Grohs (Vienna)** Phase transitions in rate-distortion theory and deep learning
10:30 – 11:00	COFFEE BREAK
11:00	**Alexandra Carpentier (Magdeburg)** Adaptive inference and its relations to sequential decision making
11:45	**Pradeep Kr. Banerjee (Leipzig)** Deep representation learning based on Blackwell sufficiency and the unique information
12:05	**Pavel Dvurechensky (Berlin)** On the complexity of optimal transport problems
12:25	CLOSING

Pradeep Kr. Banerjee (Max Planck Institute for Mathematics in the Sciences, Leipzig)

Deep representation learning based on Blackwell sufficiency and the unique information

One of the core ideas of deep learning is to obtain high level representations of data that might explain and disentangle its factors of variation. Ideally, these representations should be applicable to a variety of tasks, and hence they should be implemented based only on general assumptions. This motivates approaches based on information theory. New theory is needed in order to quantify and compute relevant notions such as the unique information of a feature or the synergistic information of a set of features in a representation. In this talk, we present a new representation learning method based on a classical ordering of information channels paradigm due to Blackwell. The fundamental quantity of interest is an information measure called the Unique Information (UI) that arises naturally in the context of a nonnegative decomposition of the mutual information into redundant and synergistic contributions. We present an alternating minimization algorithm to compute the UI and demonstrate the utility of our method for learning minimally sufficient representations in a supervised learning task. This talk summarizes joint work with Guido Montúfar (UCLA Math.) and Johannes Rauh (MPI MiS).

References:

[1] D. Blackwell, Equivalent comparisons of experiments, The Annals of Mathematical Statistics, vol. 24, no. 2, pp. 265-272, 1953.

[2] N. Bertschinger, J. Rauh, E. Olbrich, J. Jost, and N. Ay, Quantifying unique information, Entropy, vol. 16, no. 4, pp. 2161-2183, 2014.

[3] P. K. Banerjee, J. Rauh, and G. Montúfar, Computing the unique information, in Proceedings of the IEEE International Symposium on Information Theory (ISIT), 2018, pp. 141-145.

[4] P. K. Banerjee and G. Montúfar, The variational deficiency bottleneck, arXiv preprint arXiv:1810.11677, 2018.

[5] P. K. Banerjee, E. Olbrich, J. Jost, and J. Rauh, Unique informations and deficiencies, in Proceedings of the 56th Annual Allerton Conference on Communication, Control and Computing, IEEE 2018, pp. 32-38.

Helmut Bölcskei (ETH Zurich)

Fundamental limits of deep neural network learning

We develop the fundamental limits of learning in deep neural networks by characterizing what is possible if no constraints on the learning algorithm and the amount of training data are imposed. Concretely, we consider Kolmogorov-optimal approximation through deep neural networks with the guiding theme being a relation between the complexity of the function (class) to be approximated and the complexity of the approximating network in terms of connectivity and memory requirements for storing the network topology and the associated quantized weights. The theory we develop educes remarkable universality properties of deep networks. Specifically, deep networks are optimal approximants for vastly different function classes such as affine systems and Gabor systems. In addition, deep networks provide exponential approximation accuracy – i.e., the approximation error decays exponentially in the number of non-zero weights in the network – of widely different functions including the multiplication operation, polynomials, sinusoidal functions, general smooth functions, and even one-dimensional oscillatory textures and fractal functions such as the Weierstrass function, both of which do not have any known methods achieving exponential approximation accuracy. We also show that in the approximation of sufficiently smooth functions finite-width deep networks require strictly smaller connectivity than finite-depth wide networks.

Alexandra Carpentier (Otto von Guericke University Magdeburg)

Adaptive inference and its relations to sequential decision making

Adaptive inference – namely adaptive estimation and adaptive confidence statements – is particularly important in high or infinite dimensional models in statistics. Indeed whenever the dimension becomes high or infinite, it is important to adapt to the underlying structure of the problem. While adaptive estimation is often possible, it is often the case that adaptive and honest confidence sets do not exist. This is known as the adaptive inference paradox. And this has consequences in sequential decision making. In this talk, I will present some classical results of adaptive inference and discuss how they impact sequential decision making.

(joint works with Andrea Locatelli, Matthias Loeffler, Olga Klopp, and Richard Nickl)

Pavel Dvurechensky (WIAS Berlin)

On the complexity of optimal transport problems

Optimal transport (OT) distances between probability measures or histograms, including the earth mover's distance and Monge-Kantorovich-Wasserstein distance play an increasing role in different machine learning tasks, such as unsupervised learning, image retrieval, classification, domain adaptation, and training GANs. Optimal transport is especially efficient for the analysis of data with geometric structure, such as images. Nevertheless, the efficiency in practice comes with the price of costly computations. In this work we consider theoretical aspects of computational complexity of approximating optimal transport distance and barycenter, the latter being generalization of the notion of mean object for non-euclidean geometry. Namely, the main question is how many arithmetic operations is needed to approximate these objects with given accuracy. We provide algorithmic upper bounds for the complexity.

The talk is based on papers http://proceedings.mlr.press/v80/dvurechensky18a.html, http://proceedings.mlr.press/v97/kroshnin19a.html.

Patrick Gelß (FU Berlin)

Tensor-based algorithms for image classification

The interest in machine learning with tensor networks has been growing rapidly in recent years. The goal is to exploit tensor-structured basis functions in order to generate exponentially large feature spaces which are then used for supervised learning. In this talk, two different tensor approaches for quantum-inspired machine learning are presented. One is a kernel-based reformulation of the previously introduced MANDy, the other an alternating ridge regression in the tensor-train format.

Philipp Grohs (University of Vienna)

Phase transitions in rate-distortion theory and deep learning

Arnulf Jentzen (University of Münster)

High-dimensional approximation capacities of deep neural networks

In the recent years deep neural networks (DNNs) have very successfully been used in numerical simulations for a wide variety of computational problems including computer vision, image classification, speech recognition, natural language processing, computational advertisement, as well as numerical approximations of PDEs. These numerical simulations indicate in many practical relevant situations that DNNs seem to have the fundamental power to overcome the curse of dimensionality in the sense that the number of parameters in the DNN grows at most polynomially both in the dimension of the target function and in the reciprocal of the prescribed approximation precision. In this lecture we present mathematical results which prove in several situations that DNNs do indeed overcome to the curse of dimensionality.

Ilja Klebanov (Zuse Institute, Berlin)

Rigorous theory of conditional mean embeddings

Conditional mean embeddings (CME) have proven themselves a powerful tool in many machine learning applications. They allow to perform conditioning (e.g. for Bayesian inference) within the feature space (RKHS) in an efficient way by providing a relation for the kernel mean embeddings of the respective probability distributions. The theory on CMEs, however, lacks mathematical rigor and several assumptions for their applicability can be weakened. We present a rigorous mathematical theory of CMEs and demonstrate a beautiful connection to Gaussian conditioning in Hilbert spaces.

Gitta Kutyniok (TU Berlin)

Transferability of spectral graph convolutional neural networks

The success of convolutional neural networks (ConvNets) on Euclidean domains ignited an interest in recent years in extending these methods to graph structured data. This led to spectral graph convolutional neural networks, where filters are defined as elementwise multiplication in the frequency domain of a graph. Since often the dataset consists of signals defined on many different graphs, the trained ConvNet should generalize to signals on graphs unseen in the training set, showing the importance of being able to transfer filters from one graph to another. The term “transferability” then refers to the condition that a single filter/ConvNet has similar repercussions on both graphs, if two graphs describe the same phenomenon, which can be regardsd as a certain type of generalization capability. However, for a long time it was believed that spectral filters are not transferable.

In this talk we aim to debunking this common misconception, by showing that if two graphs discretize the same continuous metric space, then a spectral filter/ConvNet has approximately the same repercussion on both graphs. Our analysis even accounts for large graph perturbations as well as allows graphs to have completely different dimensions and topologies, only requiring that both graphs discretize the same underlying continuous space.

This is joint work with R. Levie, W. Huang, L. Bucci, and M. Bronstein.

Andrea Manzoni (Politecnico di Milano)

Nonlinear dimensionality reduction of parametrized PDEs using deep learning techniques

We present state-of-art projection-based reduced order models (ROMs) for parametrized PDEs, with special emphasis on the reduced basis method for nonlinear time-dependent problems. We propose new strategies to construct ROMs exploiting deep learning techniques, such as convolutional neural networks. Both the construction of a reduced-order space and the solution of the resulting reduced-order problem can be designed by exploiting neural networks for the sake of computational efficiency. Applications of interest deal with problems featuring highly nonlinear solution manifolds, as well as strong dependence of the solution on parameters, and requiring higher spatio-temporal accuracy; a relevant example is provided by coupled problems related with cardiac electrophysiology, whose goal is modeling the propagation of electric potentials in the cardiac muscle. In this case, suitable combinations of (physics-based) projection-based ROMs and (data-driven) deep/machine learning techniques can provide remarkable computational savings without affecting numerical accuracy substantially.

Ivan Oseledets (Skolkovo Institute of Science and Technology, Moscow)

Greedy algorithm for learning deep neural network architectures

In this talk I will cover our recent work on approximation of functions using Deep-ReLU type networks (https://arxiv.org/abs/1910.12686). There is a lot of results that show the expressive power of deep neural network architectures, but these papers contain very few numerical examples. I will show how we can alleviate this by learning basis functions one-by-one in a greedy manner, similar to the way basis functions are selected in compressed sensing, and confirm exponential convergence for different model problems.

Philipp Petersen (University of Vienna)

High-dimensional approximation by deep neural networks in the context of parametric PDEs

We analyse to what extent deep learning techniques are capable of efficiently solving certain parametric partial differential equations. First, we identify examples of high-dimensional parametric PDEs, where the solution map, i.e., the map from the parameter space to the solution of the PDE, can be represented by deep neural networks without a curse of dimension. Second, we present a numerical scheme, which, for these problems, trains deep network architectures to closely approximate the solution map without a noticeable curse of dimension. Finally, we analyse the number of required samples for successful training and compare this with the numbers of required snapshots for the reduced basis method.

This is joint work with Gitta Kutyniok, Mones Raslan, Reinhold Schneider, and Moritz Geist.

Johannes Rauh (Max Planck Institute for Mathematics in the Sciences, Leipzig)

Approximation properties of stochastic neural networks

Stochastic networks can be used to model joint or conditional probability distributions, both in a generative setting and in a discriminative setting, with the output reflecting preferences for different class labels given the inputs. In contrast to the case of deterministic neural networks, even if the sets of inputs and outputs are finite, the class of stochastic mappings is not. Two prominent types of stochastic networks which have played a key role in the resurgence of deep learning are Boltzmann machines and sigmoid belief networks. We give an overview of older and more recent results on the representational power of these networks, both in terms of classes of standard probabilistic graphical models that they can capture, and also in terms of the worst case and expected approximation errors that they incur depending on the number of hidden units. A focus of the talk will be on the mathematical tools used to obtain these results.

This talk summarizes joint work with Guido Montúfar and Thomas Merkh.

Leon Sallandt (TU Berlin)

A near model-free method for solving the Hamilton-Jacobi-Bellman equation in high dimensions

We treat infinite horizon optimal control problems by solving the associated stationary Hamilton-Jacobi-Bellman (HJB) equation numerically, for computing the value function and an optimal feedback area law. We use the policy iteration algorithm to reduce the highly nonlinear HJB equation to a sequence of linear equations in high dimensions. This linear equation is reformulated using the Koopman operator as an operator equation. Using the method of characteristics, we obtain a near model-free formulation of the problem. As the underlying ODE is assumed to be high-dimensional, the linearized HJB is suffering from the curse of dimensions. To overcome numerical infeasability we use low-rank hierarchical tensor product approximation, or tree-based tensor formats, in particular tensor trains (TT tensors) and multi-polynomials, since the resulting feedback law is expected to be smooth. The resulting operator equation is solved using high-dimensional quadrature, e.g. Variational Monte-Carlo methods. From the knowledge of the value function at computable samples x_i we infer the function $x \mapsto v(x)$.

(joint work with Mathias Oster and Reinhold Schneider)

Christoph Schwab (ETH Zurich)

Deep neural network expression for PDEs with uncertain input and data

For Bayesian Inverse Problems with data-to-response maps given by well-posed PDEs and subject to uncertain parametric or function space input data, under rather general conditions, we establish parametric holomorphy of the map relating observation data to the Bayesian estimate for the quantity of interest "QoI" for short). We infer expression rate bounds for this "Data-to-QoI" map by deep ReLU neural networks (and other architectures) which are uniform with respect to the data in a compact subset of R^K. We also analyze the generalization error of the DNN emulation of the Data-to-QoI map given by the Bayesian estimate, being naturally related to noisy data.

Joint work with Lukas Herrmann (ETH) and Jakob Zech (MIT).

Ingo Steinwart (University of Stuttgart)

Some thoughts and questions towards a statistical understanding of DNNs

So far, our statistical understanding of deep neural networks (DNNs) is rather limited, in particular if over-parametrized DNNs are considered. Part of the reasons for this lack of understanding is the fact that for such large DNNs the tools of classical statistical learning theory can no longer be applied. Nonetheless, over-parametrized DNNs are typically performing well in practice. In this talk, I will discuss some recent result and resulting possible questions that may be relevant for a successful end-to-end analysis of DNNs.

Edwin Stoudenmire (Flatiron Institute, New York)

Theory of a generative machine learning algorithm based on matrix product states

Matrix product states (MPS), or tensor trains, are a powerful factorization of high-order tensors. Using MPS to parameterize high-dimensional joint probability distributions gives good results for generative modeling, but perhaps of even more interest is the potential of MPS for developing detailed theories of learning. We show a first step in this direction by developing a theory of a particular MPS training algorithm which allows us to predict the fraction of a model data set (even-parity bit strings) needed to achieve a given generalization error.

Philipp Trunschke (TU Berlin)

Image classification with tensor networks

Tensor networks provide compressed representations of high-dimensional functions and are successfully used in quantum physics and uncertainty quantification. They have also been used to parameterize models in machine learning applications but suffer from the same problem as neural networks, namely that the topology of the network has to be chosen (arbitrarily) in advance. We propose a new way to look at the learning problem that allows us to develop a fully adaptive algorithm for image classification and present compelling numerical evidence of its performence.

Dmitry Yarotsky (Skolkovo Institute of Science and Technology, Moscow)

The fast non-classical approximation phase in deep neural networks

Theoretically, if approximation by a deep neural network is optimized for accuracy vs. complexity with the complexity measured by the number of network connections, one can achieve approximation rates that exceed standard classical rates known for Sobolev balls. This can be achieved by tightly encoding information about the approximated function in a small number of weights (assuming a sufficiently precise underlying arithmetic), and implementing a suitable decoder in the network. I will review various aspects of this phenomenon and its connections to earlier results on approximations and neural networks.

List of Participants

Reyhaneh Abbasi
Acoustic Research Institute, Vienna

Andreas Adelmann
Paul Scherrer Institut, Villigen & ETH

Mazen Ali
Ulm University

Pradeep Kr. Banerjee
MPI for Mathematics in the Sciences, Leipzig

Jo Andrea Brüggemann
WIAS Berlin

Helmut Bölcskei
ETH Zurich

Alexandra Carpentier
Otto von Guericke University Magdeburg

Fateme Darlik
Luxembourg University

Niklas Dexheimer
University of Mannheim

Darina Dvinskikh
WIAS Berlin

Pavel Dvurechensky
WIAS Berlin

Matthias Eckardt
HU Berlin

Matthias Ehrhardt
Bergische Universität Wuppertal

Michael Eiermann
TU Bergakademie Freiberg

Martin Eigel
WIAS Berlin

Nando Farchmin
TU Berlin

Patrick Gelß
FU Berlin

Jan Gerken
MPI for Gravitational Physics, Potsdam

Philipp Grohs
University of Vienna

Robert Gruhlke
WIAS Berlin

Martin Hess
SISSA, Trieste

Arnulf Jentzen
University of Münster

Max Kirstein
Bosch

Ilja Klebanov
Zuse Institute, Berlin

Nadja Klein
HU Berlin

Lucas Kock
HU Berlin

Gitta Kutyniok
TU Berlin

Alexander Lobbe
University of Oslo

Jan Macdonald
TU Berlin

Andrea Manzoni
Politecnico di Milano

Manuel Marschall
WIAS Berlin

Ivan Oseledets
Skolkovo Institute of Science and Technology, Moscow

Benjamin Peters
Otto von Guericke University Magdeburg

Paolo Pigato
WIAS Berlin

Lorenz Richter
FU Berlin

Leon Sallandt
TU Berlin

Christoph Schwab
ETH Zurich

Ingo Steinwart
University of Stuttgart

Claudia Strauch
University of Mannheim

Lukas Trottner
University of Mannheim

Simon Weissmann
University of Mannheim

Tizian Wenzel
University of Stuttgart

Wei Zhang
Zuse Institute, Berlin

Philipp Petersen
University of Vienna

Johannes Rauh
MPI for Mathematics in the Sciences, Leipzig

Johanna Rubisch
HTW Dresden

Reinhold Schneider
TU Berlin

Björn Sprungk
Georg-August-University Göttingen

Edwin Stoudenmire
Flatiron Institute, New York

Vikram Sunkara
FU Berlin/Zuse Institute, Berlin

Philipp Trunschke
TU Berlin

Sören Weniger
Otto von Guericke University Magdeburg

Dmitry Yarotsky
Skolkovo Institute of Science and Technology, Moscow