Workshop on Mathematics of Deep Learning

Weierstrass Institute
for Applied Analysis and Stochastics
September 13 – 15, 2017

www.wias-berlin.de/workshops/DL17/

Weierstraß-Institut für
Angewandte Analysis und Stochastik

www.wias-berlin.de

ECMath
Einstein Center
for Mathematics Berlin

Structural Inference in Statistics
Adaptation and Efficiency
FOR 1735

FOR 2402

Contents

Welcome

Dear Participant,

Welcome to the Weierstrass Institute for Applied Analysis and Stochastics in Berlin! The *Workshop on Mathematics for Deep Learning* aims to target a more analytical viewpoint on Deep Learning as is often pursued, relating it to different modern discplines of mathematics and stochastics such as high-dimensional approximation, tensor methods, sparse sampling, uncertainty quantification and probabilistic optimization. We believe that this direction will gain momentum and importance in the coming years.

Deep Learning has evolved into one of the hot topics in industry and science with a wide range of applications related to the processing and interpretation of large amounts of data. While the success and progress of recent neural network architectures has been breathtaking, the mathematical understanding and analysis of these networks is still in its infancy. However, a better understanding of the underlying structures would allow for the development of more efficient algorithms and shed light on the expressive power of architectures. Moreover, this also is a major issue for the application of deep learning methods in safety critical industrial areas such as autonomous driving.

Given below is some general information regarding logistics and other arrangements for our workshop.

Entrance to the building will be provided upon presenting your participant's badge. Please keep it on you all the time since the receptionist is supposed not to let you in without it.

Lunch can be taken at a number of restaurants and snack bars in the neighbourhood of the institute, see an extra sheet for more details.

Dinner will be held in the restaurant *UMSPANNWERK Ost*, Palisadenstr. 48, 10243 Berlin, on Thursday, September 14, 2017, at 7 p.m.

Smoking in the building is not allowed.

The workshop is jointly organized by the TU Berlin and WIAS with kind support of ECMath/MATHEON, FOR 1735 and FOR 2042.

We wish you a pleasant stay at the Weierstrass Institute and in Berlin!

Yours sincerely,

Martin Eigel, Peter Friz, Gitta Kutyniok, Reinhold Schneider, Volodia Spokoiny
(Organizers)

Wednesday, 13.09.2017

13:00	Registration & coffee
13:30	Opening
13:45	**Harald Oberhauser (Oxford)** Learning from the order of events
14:30	**Gerard Kerkyacharian (Paris)** Wavelets: A link between statistics, geometry, and probability
15:15 – 15:45	Coffee Break
15:45	**Anthony Nouy (Nantes)** Principal component analysis in tree tensor networks for high-dimensional approximation
16:30	**John Schoenmakers (Berlin)** Optimal stopping and control via approximative dynamic programming and Deep Learning

Thursday, 14.09.2017

09:00	**Ivan Oseledets (Moscow)** Review on tensor methods, Deep Learning, and some new results
10:30 – 11:00	Coffee Break
11:00	**Nadav Cohen (Jerusalem)** Expressive efficiency and inductive bias of convolutional networks: Analysis and design through hierarchical tensor decompositions
11:45	**Philipp Petersen (Berlin)** Optimal classification by deep ReLU networks
12:30 – 14:00	Lunch Break
14:00	**Stephane Mallat (Paris)** Multiscale high-dimensional learning and deep neural networks
15:30 – 16:00	Coffee Break
16:00	**Hao Ni (London)** The signature-based learning algorithm for sequential data mining and its applications
16:45	**Mike Espig (Zwickau)** An efficient method for statistical learning by means of tensor format representations
19:00	Dinner

Friday, 15.09.2017

09:00	**Thomas Wiatowski (Zürich)** Energy propagation in deep convolutional neural networks
09:45	**Dominique Picard (Paris)** Clustering high dimensional data with sparsity
10:30 – 11:00	Coffee Break
11:00	**Sergei Pereverzyev (Linz)** Regularization by the linear functional strategy in multiple kernel learning and in reduction of complexity of learning tasks
11:45	**Philipp Grohs (Vienna)** Optimal approximation with sparsely connected deep neural networks
12:30	**Martin Spindler (Hamburg)** High-dimensional L_2boosting: Rate of convergence
13:15	Closing

Nadav Cohen (The Hebrew University of Jerusalem)

Expressive efficiency and inductive bias of convolutional networks: Analysis and design through hierarchical tensor decompositions

The driving force behind convolutional networks – the most successful deep learning architecture to date, is their expressive power. Despite its wide acceptance and vast empirical evidence, formal analyses supporting this belief are scarce. The primary notions for formally reasoning about expressiveness are efficiency and inductive bias. Expressive efficiency refers to the ability of a network architecture to realize functions that require an alternative architecture to be much larger. Inductive bias refers to the prioritization of some functions over others given prior knowledge regarding a task at hand. Through an equivalence to hierarchical tensor decompositions, we study the expressive efficiency and inductive bias of various convolutional network architectural features. Our results shed light on the demonstrated effectiveness of convolutional networks, and in addition, provide new tools for network design. The talk is based on a series of works from COLT'16, ICML'16, CVPR'16 and ICLR'17 (as well as several new pre-prints), with collaborators Or Sharir, Yoav Levine, Ronen Tamari, David Yakira and Amnon Shashua.

Mike Espig (Westsächsische Hochschule Zwickau (WHZ))

An efficient method for statistical learning by means of tensor format representations

The coming century is surely the century of high dimensional data. With the rapid growth of computational methods in all areas of economy and science, high-dimensional data becomes very common. Thus, analyzing high-dimensional data is an urgent problem of great practical importance. However, there are some unique challenges for analyzing data of high dimensions including the curse of dimensionality and the meaningful gaining of knowledge, or learning, from high-dimensional datasets. With standard techniques it is impossible to store all entries of the high-dimensional data explicitly. The reason is that the computational complexity and the storage cost are growing exponentially with the number of dimensions. Besides of the storage one should also solve this high-dimensional problems in a reasonable (e.g. linear) time and obtain a solution in some compressed (low-rank/sparse) tensor formats. The complexity of many existing data analyzing algorithms is exponential with respect to the number of dimensions. With increasing dimensionality, these algorithms soon become computationally intractable and therefore inapplicable in many real applications. During the last years, tensor format representation techniques were successfully applied to high-dimensional problems. In the current work, tensor format representation technics are introduced to the statistical learning problem.

Philipp Grohs (University of Vienna)

Optimal approximation with sparsely connected deep neural networks

We derive fundamental lower bounds on the connectivity and the memory requirements of deep neural networks guaranteeing uniform approximation rates for arbitrary function classes in $L^2(\mathbb{R}^d)$. In other words, we establish a connection between the complexity of a function class and the complexity of deep neural networks approximating functions from this class to within a prescribed accuracy. Additionally, we prove that our lower bounds are achievable for a broad family of function classes. Specifically, all function classes that are optimally approximated by a general class of representation systems – so-called *affine systems* – can be approximated by deep neural networks with minimal connectivity and memory requirements. Affine systems encompass a wealth of representation systems from applied harmonic analysis such as wavelets, ridgelets, curvelets, shearlets, α-shearlets, and more generally α-molecules. This result elucidates a remarkable universality property of neural networks and shows that they achieve the optimum approximation properties of all affine systems combined. As a specific example, we consider the class of $1/\alpha$-cartoon-like functions, which is approximated optimally by α-shearlets. We also explain how our results can be extended to the case of functions on low-dimensional immersed manifolds. Finally, we present numerical experiments demonstrating that the standard stochastic gradient descent algorithm generates deep neural networks providing close-to-optimal approximation rates at minimal connectivity. Moreover, these results show that stochastic gradient descent actually learns approximations that are sparse in the representation systems optimally sparsifying the function class the network is trained on.

Joint work with Helmut Bölcskei, Gitta Kutyniok and Philipp Petersen.

Gerard Kerkyacharian (LPMA-CREST, Paris)

Wavelets: A link between statistics, geometry, and probability

Wavelet theory was developed by Meyer, Daubechies, Lemarie, Mallat, Cohen and scores of other mathematicians more than thirty years ago, after the work of Frazier, Jawerth, and Weiss. Lots of applications have been made thereafter. For example, the Littlewood-Paley analysis and wavelet theory have proved to be a very useful tool in nonparametric statistic analysis. This is essentially due to the fact that most of the regularity (Sobolev and Besov) spaces can be characterized by wavelet sparse coefficients. In turn, in the nineties the wavelet theory allowed to develop ([3]) an adaptive estimator of the density of a probability law with no apriori knowledge of the regularity. Then it appeared that the Euclidian analysis is not always appropriate because many statistical problems have their own geometry. For instance, this is the situation in Tomography, where one uses Harmonic analysis of the ball, and in the study of the Cosmological Microwave Background, which requires Harmonic analysis on the sphere ([2]). At the same time the wavelet theory was extended in various geometric and nonclassical frameworks. Extensions of this kind have already been implemented in the cases of the interval ([9]), the ball ([10]), the sphere ([7, 8]), and have been extensively used in statistical applications (see for instance ([4]). In recent years the Littlewood-Paley analysis and wavelet theory were developed in the general framework of Riemannian manifolds and furthermore in the general setting of a positive operator associated to a suitable Dirichlet space with a good behavior of the associated heat kernel ([1, 5]). Among other things this theory allows to revisit the old problem of almost everywhere regularity of Gaussian fields ([6]). In this talk we will review the topics mentioned above and present some new results.

References

[1] T. Coulhon, G. Kerkyacharian, P. Petrushev, Heat kernel generated frames in the setting of Dirichlet spaces. J. Fourier Anal. Appl. 18(5) (2012), 995–1066.

[2] P. Baldi, G. Kerkyacharian, D. Marinucci, D. Picard, Asymptotic for spherical needlets. Annals of Statistics, Vol 37, No. 3 (2009), 1150–1171.

[3] D. Donoho, I. Johnstone, G. Kerkyacharian, D. Picard, Wavelet shrinkage: Asymptotia. Journal of the Royal Statistical Society as “Special Read Paper”, 57, No. 2, (1995), 301–369.

[4] G. Kerkyacharian, G. Kyriazis, E. Le Pennec, P. Petrushev, D. Picard, Inversion of noisy radon transform by svd based needlet. Appl. Comput. Harmon. Anal. 28 (2010), 24–45.

[5] G. Kerkyacharian, P. Petrushev, Heat kernel based decomposition of spaces of distributions in the framework of Dirichlet spaces. Trans. Amer. Math. Soc. 367 (2015), 121–189.

[6] G. Kerkyacharian, P. Petrushev, D. Picard, S. Ogawa, Regularity of Gaussian processes on Dirichlet spaces. arXiv:1508.00822 (4 August 2015).

[7] F. Narcowich, P. Petrushev, J. Ward, Local tight frames on spheres. SIAM J. Math. Anal. 38 (2006), 574–594.

[8] F. J. Narcowich, P. Petrushev, J. Ward, Decomposition of Besov and Triebel-Lizorkin spaces on the sphere. J. Funct. Anal. 238 (2006), 530–564.

[9] P. Petrushev, Yuan Xu, Localized polynomial frames on the interval with Jacobi weights. J. Fourier Anal. Appl. 11(5) (2005), 557–575.

[10] P. Petrushev, Yuan Xu, Localized polynomials frames on the ball. Constr. Approx. 27 (2008), 121–148.

[S] L. Saloff-Coste, *Aspects of Sobolev-type inequalities*, volume 289 of *London Mathematical Society Lecture Note Series*. Cambridge University Press, Cambridge, 2002.

Stephane Mallat (Ecole Normale Supérieure, Paris)

Multiscale high-dimensional learning and deep neural networks

Classifications and regressions from data requires to approximate functions in high dimensional spaces. Deep convolutional networks seem perform such approximations while avoiding the curse of dimensionality, by computing invariants which are learned from data. It raises issues in many branches of mathematics including statistics, probability, harmonic analysis and geometry. We begin by reviewing two-layer neural networks and their approximation limitations. We analyze deeper neural network architectures in relation with multiscale wavelet decompositions, and the calculation of stable invariants. We consider unsupervised and supervised learning problems. Applications are shown for image and audio classification, for quantum energy regression, and for modelization of complex random processes including textures and turbulences in statistical physics.

Hao Ni (University College London/The Alan Turing Institute)

The signature-based learning algorithm for sequential data mining and its applications

Regression analysis aims to use observational data from multiple observations to develop a functional relationship relating explanatory variables to response variables, which is important for much of modern statistics, and econometrics, and also the field of machine learning. In this talk, we consider the special case where the explanatory variable is a stream of information, and the response is also potentially a stream. We provide an approach based on identifying carefully chosen features of the stream which allows linear regression to be used to characterise the functional relationship between explanatory variables and the conditional distribution of the response; the methods used to develop and justify this approach, such as the signature of a stream and the shuffle product of tensors, are standard tools in the theory of rough paths and seem appropriate in this context of regression as well and provide a surprisingly unified and non-parametric approach. We believe that the insight provided by this paper will provide an additional tool in the toolbox for studying sequential data.

Moreover, we apply our method to some of the empirical datasets, including online Chinese handwriting character data and action classification, which achieve the state-of-art recognition results.

Anthony Nouy (Ecole Centrale de Nantes)

Principal component analysis in tree tensor networks for high-dimensional approximation

We present an extension of principal component analysis for functions of multiple random variables and an associated algorithm for the approximation of such functions using tree-based low-rank formats (tree tensor networks). A multivariate function is here considered as an element of a Hilbert tensor space of functions defined on a product set equipped with a probability measure, the function being identified with a multidimensional array when the product set is finite. The algorithm only requires evaluations of functions (or arrays) on a structured set of points (or entries) which is constructed adaptively. The algorithm constructs a hierarchy of subspaces associated with the different nodes of a dimension partition tree and a corresponding hierarchy of interpolation operators. Optimal subspaces are estimated using empirical principal component analysis of interpolations of partial random evaluations of the function. The algorithm is able to provide an approximation in any tree-based format with either a prescribed rank or a prescribed relative error, with a number of evaluations of the order of the storage complexity of the approximation format.

Harald Oberhauser (University of Oxford)

Learning from the order of events

Learning from path-valued data can be a challenging topic due to the infinite dimensionality and non-local compactness of the space of paths. I will talk about how recent ideas from stochastic analysis and rough path theory can give rise to nonparametric methods; vice versa, approaches from the machine learning community can yield new proofs and extensions of results in rough path theory and stochastic analysis.

Joint works with Ilya Chevyrev, Franz Kiraly, Terry Lyons.

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

Review on tensor methods, Deep Learning, and some new results

Tensor network is a compact way to represent multivariate functions; deep neural networks can be also thought as a compact way to represent multivariate functions. There are many other direct or indirect connections between these two fields. Recent work by Cohen, Shashua, and others elaborate on this connection by interpreting a tensor decomposition as a special case of a neural network. This does not directly lead to new algorithms, but already provides insight. In this talk, I will highlight several interesting recent results for using tensors and deep learning.

Sergei Pereverzyev (The Johann Radon Institute for Computational and Applied Mathematics (RICAM), Linz)

Regularization by the linear functional strategy in multiple kernel learning and in reduction of complexity of learning tasks

The choice of the kernel is known to be a challenging and central problem of kernel based supervised learning. Recent applications and significant amount of literature have shown that using multiple kernels (the so-called Multiple Kernel Learning (MKL)) instead of a single one can enhance the interpretability of the learned function and improve performances. However, a comparison of existing MKL-algorithms shows that though there may not be large differences in terms of accuracy, there is difference between MKL-algorithms in complexity as given by the training time, for example. In this talk we present a promising approach for training the MKL-machine by the linear functional strategy, which is either faster or more accurate than previously known ones. Moreover, we also briefly discuss a possibility of combining our MKL-strategy with a dictionary of “deep kernelsäppearing at the top of a trained convolutional network.

Philipp Petersen (TU Berlin)

Optimal classification by deep ReLU networks

In this talk, we analyze function approximation by deep neural networks. In particular, we are interested in the approximation of functions that assume only finitely many values, i.e. piecewise constant functions. This analysis is motivated by the recent successful applications of deep neural networks in classification tasks, where complex and high-dimensional data is mapped to only a few classifying labels. We restrict ourselves to the ReLU activation function, which is probably the most widely used in applications. For this activation function, we will demonstrate that deep neural networks yield optimal approximation rates in an information theoretical sense. Using a concrete construction, we show that, for $n, d \in \mathbb{N}$ arbitrary, the number of edges and the depth of a ReLU neural network approximating a piecewise C^n function on $\mathbb{R}^d$ with jump singularities along C^n curves, scale optimally with respect to the approximation quality. To study the optimality, we establish the asymptotic description complexity of piecewise constant functions and interpret networks as encoder-decoder pairs. Fundamental limits on the efficiency of encoding algorithms then provide lower bounds of neural networks.

This talk is based on joint work with Felix Voigtlaender.

Dominique Picard (University Paris-Diderot)

Clustering high dimensional data with sparsity

We consider the problem of clustering high dimensional data. We observe a matrix Y of size $n \times d$. Typically d is much larger than n (but not necessarily) and each column vector represents an individual denoted by $Y_i,\ i \leq n$, of dimension d.

For sake of simplicity, we assume that there is only two classes, i.e. that there exists $A \subset \{1, \dots n\}$ and two vectors of $\mathbb{R}^d$, θ_- and θ_+, such that $Y_i \sim N(\theta_-, \sigma^2 I)$ for $i \in A$, $Y_i \sim N(\theta_+, \sigma^2 I)$ for $i \in A^c$.

We assume that θ_- and θ_+ are unknown and sparse in the sense that they belong to the regularity set:

$$\Theta(s, L) := \left\{ \theta \in \mathbb{R}^d,\ \sup_{K \in \mathbb{N}^*} K^{2s} \sum_{k \geq K} (\theta^k)^2 \leq L^2 \right\}.$$

Again, for sake of simplicity we suppose that A is of the form $\{1, \dots, n\tau\}$, for $\tau \in (0, 1)$ unknown. We also assume that in fact that $\tau \in]\varepsilon, 1 - \varepsilon[$ (for some unknown and fixed parameter $0 < \varepsilon < 1/2$) and we put:

$$\Delta^2 = \sum_{l=1}^{d} (\theta_+ - \theta_-)^2.$$

Our problem is to determine whether or not it is efficient to smooth the data i.e. to replace the vectors $Y_i := Y_i(d), i \leq n$ by, for $T < d$, $Y_i(T), i \leq n$, the vectors of $\mathbb{R}^T$, of the T first coordinates of Y_i.

Then, if smoothing reveals to be useful, how to choose T ideally in an adaptive way (without knowing the regularity s).

We also propose an adaptive algorithm to estimate θ_+ and θ_- with minimax rates.

1. Smoothing rates

We consider the following family of algorithms.

$$\hat{\tau}(T) = \frac{1}{n} ArgMin_{k \in \{2, \dots, n-2\}} \left\{ \sum_{j \leq k} \sum_{\ell \leq T} \left(Y_j^\ell - \frac{1}{k} \sum_{j \leq k} Y_j^\ell \right)^2 + \sum_{j \geq k+1} \sum_{\ell \leq T} \left(Y_j^\ell - \frac{1}{n-k} \sum_{j \geq k+1} Y_j^\ell \right)^2 \right\}.$$

Proposition 1. *If we assume that there exists a constant* R

$$\Delta^2 \geq R \Big[T^{-2s} \vee \frac{\sigma^2 T}{n} \Big]$$

then there exists constants $c_2,\ c_3$, *and* κ *such that,*

$$P\Big(\hat{\tau}(T) - \tau \geq \kappa \frac{\sigma^2 T}{n\Delta^2} \Big) \leq 2n[\exp\{-c_2 RT\} + \exp\{-c_3 \kappa T\}].$$

Optimum is obtained for

$$T_s := \left[\frac{n}{\sigma^2} \right]^{\frac{1}{1+2s}}.$$

2. Adaptative choice for T

Form the following pseudo-data in $\mathbb{R}^d$: Z

$$Z^\ell = \frac{1}{n}\sum_{j=1}^{n} Y_j^\ell - \frac{2}{n}\sum_{j=1}^{n/2} Y_j^\ell, \ell = 1, \dots, d.$$

Consider the Lepski smoothers (c is a tuning constant)

$$\hat{T} := \min\left\{k,\ \sum_{m=k'}^{l} [Z^m]^2 \leq cl\frac{\sigma^2}{n}\log[d \vee n],\ \forall d \geq l \geq k' \geq k\right\}.$$

Theorem 1. *We assume that θ_+ and θ_- are in $\Theta(s, L)$. We suppose that there exists a constant $a > 0$ such that*

$$\frac{n}{\sigma^2} \geq a \log d.$$

Then, if there exists a constant $R = R(L, \varepsilon)$ such that

$$\Delta^2 \geq R\Big[\frac{\sigma^2 \log[d \vee n]}{n}\Big]^{\frac{2s}{1+2s}}, \tag{1}$$

then for any γ, and for c large enough,

$$P\left(\hat{\tau}(\hat{T}) - \tau \geq \kappa\Big[\frac{\sigma^2 \log[d \vee n]}{n}\Big]^{\frac{2s}{1+2s}}\Delta^{-2}\right) \leq [d \vee n]^{-\gamma}. \tag{2}$$

3. Adaptation rates for θ_- and θ_+, case $\sigma^2 = \frac{\sigma_0^2}{d}$

We first detect the change using the procedure above, using $\hat{T}$, and then $\hat{\tau} = \hat{\tau}(\hat{T})$.

Then we estimate θ_- and θ_+, with the following procedure. (We denote $\theta_\pm$ for respectively either θ_- or θ_+:

$$\hat{\theta}_\pm = (\hat{\theta}_\pm^1, \dots, \hat{\theta}_\pm^d),\ \hat{\theta}_\pm^k := \hat{\theta}_{\pm,k} I\{k \leq \hat{T}^*(\pm)\}$$

$$\hat{\theta}_{-,k} := \frac{1}{n\hat{\tau}(\hat{T})}\sum_{j=1}^{n\hat{\tau}(\hat{T})} Y_j^k \quad \hat{\theta}_{+,k} := \frac{1}{n(1-\hat{\tau}(\hat{T}))}\sum_{j=n\hat{\tau}(\hat{T})+1}^{n} Y_j^k$$

$$\hat{T}^*(\pm) := \min\left\{k,\ \sum_{m=k+1}^{l} [\hat{\theta}_{\pm,m}]^2 \leq cl\frac{\sigma^2}{n}\log[d \vee n],\ \forall l \geq k+2\right\}.$$

And we prove the following result meaning that (without condition on Δ this time) we are able to adaptively estimate θ_- and θ_+ with minimax rates up to logarithmic factors.

Theorem 2. *With the estimates defined above, then, for $s > 0$, $c > c_0$, we have that there exists a constant C such that:*

$$\sup_{\theta_\pm \in \Theta(s,L)} E\|\hat{\theta}_\pm - \theta_\pm\|_2^2 \leq C\left\{\frac{nd}{\log[n \vee d]}\right\}^{\frac{-2s}{1+2s}}. \tag{3}$$

John Schoenmakers (Weierstrass Institute (WIAS), Berlin)

Optimal stopping and control via approximative dynamic programming and Deep Learning

We consider application of deep neural networks for solving multi-dimensional optimal stopping and control problems that can be solved via backward dynamic program principles. Also we propose and discuss the combination of linear regression methods with ideas of multi-layer deep neural networks for solving such kind of problems.

Martin Spindler (University of Hamburg)

High-dimensional L_2boosting: Rate of convergence

Boosting is one of the most significant developments in machine learning. This paper studies the rate of convergence of L_2Boosting, which is tailored for regression, in a high-dimensional setting. Moreover, we introduce so-called " post-Boosting". This is a post-selection estimator which applies ordinary least squares to the variables selected in the first stage by L_2Boosting. Another variant is "Orthogonal Boosting" where after each step an orthogonal projection is conducted. We show that both post-L_2Boosting and the orthogonal boosting achieve the same rate of convergence as LASSO in a sparse, high-dimensional setting. We show that the rate of convergence of the classical L_2Boosting depends on the design matrix described by a sparse eigenvalue constant. To show the latter results, we derive new approximation results for the pure greedy algorithm, based on analyzing the revisiting behavior of L_2Boosting. We also introduce feasible rules for early stopping, which can be easily implemented and used in applied work. Our results also allow a direct comparison between LASSO and boosting which has been missing from the literature. Finally, we present simulation studies and applications to illustrate the relevance of our theoretical results and to provide insights into the practical aspects of boosting. In these simulation studies, post-L_2Boosting clearly outperforms LASSO.

Thomas Wiatowski (ETH Zurich)

Energy propagation in deep convolutional neural networks

Deep convolutional neural networks (CNNs) used in practice employ potentially hundreds of layers and 10,000s of nodes. Such network sizes entail significant computational complexity due to the large number of convolutions that need to be carried out; in addition, a large number of parameters needs to be learned and stored. Very deep and wide CNNs may therefore not be well suited to applications operating under severe resource constraints as is the case, e.g., in low-power embedded and mobile platforms. In this talk, we aim at understanding the impact of CNN depth on the network's feature extraction capabilities. Specifically, we analyze how many layers are actually needed to have "most" of the input signal's features be contained in the feature vector generated by the network. This question can be formalized by asking how quickly the energy contained in the propagated signals (a.k.a. feature maps) decays across layers. We address this question for the class of scattering networks that employ general filters, the modulus non-linearity, and no pooling, and find that under mild analyticity and high-pass conditions on the filters (which encompass, inter alia, various constructions of Weyl-Heisenberg filters, wavelets, ridgelets, (α)-curvelets, and shearlets) the feature map energy decays at least polynomially fast. For broad families of wavelets and Weyl-Heisenberg filters, the guaranteed decay rate is shown to be exponential. Our results yield handy estimates of the number of layers needed to have at least $((1-\varepsilon)\cdot 100)\%$ of the input signal energy be contained in the feature vector. Finally, we show how networks of fixed (possibly small) depth can be designed to guarantee that most of the input signal's energy are contained in the feature vector.

The talk represents joint work with H. Bölcskei and P. Grohs.

List of Participants

Randolf Altmeyer
HU Berlin

Gilles Blanchard
University of Potsdam

Nadav Cohen
The Hebrew University
of Jerusalem

Martin Eigel
WIAS Berlin

Christian Etmann
University of Bremen

Mohamed Gaafar
TU Berlin

Philipp Grohs
University of Vienna

Franziska Göbel
University of Potsdam

Lukas Herrmann
ETH Zurich

Sadegh Jokar
GameDuell GmbH, Berlin

Gerard Kerkyacharian
LPMA-CREST, Paris

Thomas Koprucki
WIAS Berlin

Gitta Kutyniok
TU Berlin

Sebastian Lunz
University of Cambridge

Stephane Mallat
Ecole Normale Supérieure, Paris

Mahdi Barzegar Khalilsarai
TU Berlin

Giuseppe Caire
TU Berlin

Simon Diehl
AVM Audiovisuelles Marketing
Computersysteme GmbH, Berlin

Mike Espig
Westsächsische Hochschule Zwickau (WHZ)

Peter Friz
TU Berlin/WIAS Berlin

Martin Genzel
TU Berlin

Robert Gruhlke
WIAS Berlin

Ali Hashemi
TU Berlin

Sahar Iravani
Zuse Institute Berlin

Aditya Kela
University of Cologne

Olaf Klein
WIAS Berlin

Sebastian Krämer
RWTH Aachen

Steffen Limmer
TU Berlin

Christian Löbbert
RWTH Aachen

Anieza Maltsi
WIAS Berlin

Maximilian März
TU Berlin

Hao Ni
University College London/
The Alan Turing Institute

Harald Oberhauser
University of Oxford

Sergei Pereverzyev
RICAM, Linz

Max Pfeffer
TU Berlin

Rafael Reisenhofer
University of Bremen

Reinhard Schachtner
Infineon Technologies AG Regensburg

John Schoenmakers
WIAS Berlin

Alexander Sikorski
Zuse Institute Berlin

Vladimir Spokoiny
WIAS Berlin

Rajesh Tamang
University of Osnabrück

Thomas Wiatowski
ETH Zurich

Sergiy Nesenenko
TU Berlin

Anthony Nouy
Ecole Centrale de Nantes

Ivan Oseledets
Skolkovo Institute of Science and Technology, Moscow

Philipp Petersen
TU Berlin

Dominique Picard
University Paris-Diderot

Markus Reiß
HU Berlin

Reinhold Schneider
TU Berlin

Christoph Schwab
ETH Zurich

Martin Spindler
University of Hamburg

Tim Sullivan
FU Berlin and Zuse Institute Berlin

Felix Voigtlaender
TU Berlin