Mathematisches Forschungsinstitut Oberwolfach

Report No. 36/2021

DOI: 10.4171/OWR/2021/36

Computation and Learning in High Dimensions (hybrid meeting)

Organized by
Albert Cohen, Paris
Wolfgang Dahmen, Columbia
Ronald A. DeVore, College Station
Angela Kunoth, Köln

1 August – 7 August 2021

Abstract. The most challenging problems in science often involve the learning and accurate computation of high dimensional functions. High-dimensionality is a typical feature for a multitude of problems in various areas of science. The so-called *curse of dimensionality* typically negates the use of traditional numerical techniques for the solution of high-dimensional problems. Instead, novel theoretical and computational approaches need to be developed to make them tractable and to capture fine resolutions and relevant features. Paradoxically, increasing computational power may even serve to heighten this demand, since the wealth of new computational data itself becomes a major obstruction. Extracting essential information from complex problem-inherent structures and developing rigorous models to quantify the quality of information in a high-dimensional setting pose challenging tasks from both theoretical and numerical perspective. This has led to the emergence of several new computational methodologies, accounting for the fact that by now well understood methods drawing on spatial localization and mesh-refinement are in their original form no longer viable. Common to these approaches is the nonlinearity of the solution method. For certain problem classes, these methods have drastically advanced the frontiers of computability. The most visible of these new methods is *deep learning*. Although the use of deep neural networks has been extremely successful in certain application areas, their mathematical understanding is far from complete.

This workshop proposed to deepen the understanding of the underlying mathematical concepts that drive this new evolution of computational methods and to promote the exchange of ideas emerging in various disciplines about how to treat multiscale and high-dimensional problems.

Mathematics Subject Classification (2010): 65-XX, 41-XX.

Introduction by the Organizers

Complex scientific models like climate models, turbulence, fluid structure interaction, nanosciences and reliability control, demand finer and finer resolution in order to increase their reliability. This demand is not simply solved by increasing computational power. Indeed, higher computability even contributes to the problem by generating wealthy data sets for which efficient organization principles are not available. Extracting essential information from complex structures and developing rigorous models for quantifying the quality of information is an increasingly important issue. These tasks become even more demanding when the problem is high dimensional, in the sense that it involves a large number of variables or parameters.

Inherently high-dimensional problems appear naturally in various scientific disciplines. Prominent examples of such problems are: (i) PDEs that describe complex processes in computational chemistry and physics, such as the Fokker-Planck and the Schrödinger equations, (ii) stochastic or parameter-dependent PDEs used in simulation and optimal control and design, (iii) classification and regression problems arising in big-data analysis with large number of input/output variables. While significant advances have been made in "forward problems" trying to exploit sparsity in effectively recovering high-dimensional functions, corresponding inverse problems like state- or parameter estimation pose even greater challenges. One reason is that one usually has to cope with a strong undersampling — a small-data problem — due to prohibitive cost or severe obstructions to acquiring observation data. An important issue is to properly formulate corresponding *data-assimilation* frameworks and to understand the role of model reduction and sparse recovery in this context.

The mathematical methods emerging to address these problems try to exploit in a subtle way the structure of the problem in order to extract the necessary information. They have several common features including the determination of whether the underlying objects have a sufficiently small information content to be computationally tractable, and how this content might be accessible through certain sparse representations. The numerical methods themselves are typically highly nonlinear with the ability of separating solution characteristics living on different length scales. Having to deal with the appearance and interaction of local features at different levels of resolution has, for instance, brought about spatially adaptive methods as a key methodology that has advanced the frontiers of computability for certain problem classes in numerical analysis. The current state of signal processing, learning theory, and numerical computation can be viewed as an evolution from the introduction of multiscale and adaptive methods to the current high dimensional methods based on concepts such as sparsity, anisotropy, model reduction, low-rank tensor methods, random projections and neural networks.

Multiscale techniques, such as wavelet decompositions, were introduced to manage the interaction of different length scales. In the very spirit of harmonic analysis they allow one to decompose complex objects into simple building blocks that again support analyzing multiscale features. Our first Oberwolfach Workshops "Wavelet and Multiscale Methods" held in July 2004 and August 2007 served to bring together the main developers of multiscale decompositions for signal processing with those using these techniques for numerical methods for PDEs and thus contributed to the growth of both of these disciplines. While multiscale techniques were first exploited primarily for treating *explicitly* given objects, like digital signals and images or data sets, the use of such concepts proved important for recovering also *implicitly* given objects, like solutions of partial differential or boundary integral equations, as well. The close marriage of discretization, analysis and the solution process based on *adaptive* wavelet methods has led to significant theoretical advances as well as new algorithmic paradigms for linear and nonlinear stationary variational problems. Through thresholding, best N-term approximation, and adaptivity, multiscale techniques from nonlinear approximation theory and harmonic analysis become practically manageable. They now are a major component of modern signal processing and modern numerical computation.

Our last three workshops in August 2010 and "Multiscale and High-Dimensional Problems held in July/August 2013 and March 2017 recognized the increasing demand on finding numerical techniques which apply to high dimensional problems. They brought together various disciplines where such problems are encountered. Those workshops not only accelerated the advancement of nonlinear and multiscale methodologies but also provided beneficial cross-fertilizations between the various areas represented in the workshop, see the Oberwolfach Reports 34/2004, 36/2007, 33/2010, 39/2013, 17/2017. Among the several recognizable outcomes of the workshops were: (i) the emergence of compressed sensing as an exciting alternative to the traditional sensing-compression paradigm, (ii) fast online computational algorithms based on adaptive partition for mathematical learning, (iii) clarification of the role of coarsening in adaptive numerical methods for PDEs, (iv) injection of the notion of sparsity into stochastic models to identify computational paradigms that are more efficient than Monte Carlo techniques, (v) a coherent theory to explain why techniques like sparse representation and reduced modeling work and how they can be improved.

This latest workshop has once again been directed at multi-scale and high dimensional problems incorporating the new emerging aspects mentioned above. It focussed on the interaction of scientists from different disciplines and thereby resulted in more rapid developments of new methodologies in these various domains. It was also a bridge from theoretical foundations to applications, such as mechanical engineering, mathematical biology, quantum chemistry, signal and image processing, complex fluid flows. Examples of conceptual issues that were addressed in our workshop were:

- adaptive and nonlinear multilevel methods for high-dimensional PDEs, for parametric PDEs and PDEs with stochastic data;
- multilevel and high-dimensional meshless methods;
- incorporating anisotropy in analysis, estimation, compression and encoding;

- interaction of different scales and variables under relevant linear and nonlinear mappings;
- convergence theory and analysis for model reduction and low-rank methods;
- numerical aspects of compressed sensing;
- design and analysis of estimators in high dimensional machine learning;
- solution concepts for problems of high spatial dimension utilizing anisotropy;
- data assimilation and inversion concepts in high dimensional settings;
- tensor structures and tensor sparsity for high dimensional approximation problems;
- identifying and analyzing model classes for wich deep neural networks perform well:
- Design and efficacy of learning algorithms

In summary, the conceptual similarities that occur in a variety of application domains suggested that a wealth of synergies and cross–fertilization should be exploited. These concepts are in our opinion not only relevant for the development of efficient solution methods for large scale and inherently high-dimensional problems but also for the formulation of rigorous mathematical models for quantifying the extraction of essential information from complex objects in many dimensions.

As in the previous workshops, the proposed participants were experts in areas like nonlinear approximation theory, statistical learning theory,compressed sensing, tensor approximations, hyperbolic cross approximation, finite elements, spectral methods, harmonic analysis and wavelets, numerical fluid mechanics, inverse problems, stochastic PDEs, PDE-constrained control problems, or model reduction.

Workshop(hybrid meeting): Computation and Learning in High Dimensions

Table of Contents

Abstracts

The Impact of Artificial Intelligence on Parametric Partial Differential Equations

Gitta Kutyniok
(joint work with Moritz Geist, Philipp Petersen, Mones Raslan,
Reinhold Schneider)

High-dimensional parametric partial differential equations (PDEs) appear in various contexts including control and optimization problems, inverse problems, risk assessment, and uncertainty quantification. In most such scenarios the set of all admissible solutions associated with the parameter space is inherently low dimensional. This fact forms the foundation for the so-called reduced basis method.

Recently, numerical experiments demonstrated the remarkable efficiency of using deep neural networks to solve parametric problems. In this talk, after an introduction into deep learning, we will present a theoretical justification for this class of approaches. More precisely, we will derive upper bounds on the complexity of ReLU neural networks approximating the solution maps of parametric PDEs. In fact, without any knowledge of its concrete shape, we use the inherent low-dimensionality of the solution manifold to obtain approximation rates which are significantly superior to those provided by classical approximation results. We use this low-dimensionality to guarantee the existence of a reduced basis. Then, for a large variety of parametric PDEs, we construct neural networks that yield approximations of the parametric maps not suffering from a curse of dimensionality and essentially only depending on the size of the reduced basis.

Finally, we present a comprehensive numerical study of the effect of approximation-theoretical results for neural networks on practical learning problems in the context of parametric partial differential equations. These experiments strongly support the hypothesis that approximation-theoretical effects heavily influence the practical behavior of learning problems in numerical analysis.

References

[1] M. Geist, P. Petersen, M. Raslan, R. Schneider, and G. Kutyniok. *Numerical Solution of the Parametric Diffusion Equation by Deep Neural Networks.* J. Sci. Comput. **88** (2021), Article number: 22.

[2] G. Kutyniok, P. Petersen, M. Raslan, and R. Schneider. *A Theoretical Analysis of Deep Neural Networks and Parametric PDEs.* Constr. Approx., to appear (arXiv:1904.00377).

Gradient Descent for learning Linear Neural Networks: Convergence, Riemannian Geometry and Implicit Bias

Holger Rauhut

Many recent breakthroughs in applications of machine learning (such as face recognition, autonomous driving, drug design and machine translation) are based on learning deep neural networks from training examples. Despite these empirical successes the mathematical understanding of the inner workings of deep learning is still in its infancy.

One usually adapts deep neural networks by minimizing a non-convex loss functional via (stochastic) gradient methods. Convergence properties are not yet well-understood in this setting. Moreover, a puzzling observation is that learning neural networks with a number of parameters exceeding the number of training examples leads to networks that generalize very well to unseen data although intuition from classical statistics would rather predict a scenario of overfitting [9]. A current working hypothesis is that the chosen optimization algorithm has a significant influence on the selection of the learned network. In fact, in this overparameterized context there are many global minimizers so that the optimization method induces an implicit bias on the computed solution. It seems that gradient descent methods and their stochastic variants favor networks of low complexity (in a suitable sense), and, hence, appear to be very well suited for large classes of real data.

Initial attempts in understanding these phenomena consider the simplified setting of linear networks, i.e., (deep) factorizations of matrices and revealed a surprising relation to the field of low rank matrix recovery in the sense that gradient descent favors low rank matrices in certain situations.

Convergence properties of learning linear networks have been studied, e.g., in [1, 2, 3, 7]. Given pairs $(x_i, y_i) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$, $i = 1, \ldots, m$ of input and output data, one considers learning a linear neural network of the form $W = W_N \cdots W_1$, i.e., a linear function in factorized form with $W_i \in \mathbb{R}^{d_i \times d_{i-1}}$, $d_0 = d_x, d_N = d_y$, such that the ℓ_2-loss

$$L^N(W_1, \ldots, W_N) = L^1(W_N \cdots W_1) = \frac{1}{2}\sum_{j=1}^{m} \|x_i - W y_i\|_2^2 = \frac{1}{2}\|X - W_N \cdots W_1 Y\|_F^2$$

is minimized. Gradient descent starts with some $\overrightarrow{W}(0) = (W_1(0), \ldots, W_N(0))$ and computes the iterates

$$\overrightarrow{W}(k+1) = \overrightarrow{W}(k) + \eta_k \nabla L^N(\overrightarrow{W}(k)). \tag{1}$$

From a mathematical view point it is also useful to analyze the corresponding gradient flow

$$\frac{d}{dt}\overrightarrow{W}(t) = -\nabla L^N(\overrightarrow{W}(t)). \tag{2}$$

As shown in [1] the quantity $W_{j+1}(0)^T W_{j+1}(0) - W_j(0) W_j(0)^T$ of the flow is constant with respect to time t, In particular, if the initial condition is balanced, i.e.,

$$W_{j+1}(0)^T W_{j+1}(0) = W_j(0) W_j(0)^T, \quad j = 1, \ldots, N-1,$$

then the flow is balanced for all times $t \geq 0$, i.e., $W_{j+1}(t)^T W_{j+1}(t) = W_j(t) W_j(t)^T$. In this case, the product $W(t) = W_N(t) \cdots W_1(t)$ satisfies the equation

$$\frac{d}{dt} W = -\sum_{j=1}^{N} (WW^T)^{\frac{N-j}{N}} \cdot \nabla_W L^1(W) \cdot (W^T W)^{\frac{j-1}{N}}. \tag{3}$$

Let $\mathcal{M}_r$ be the manifold or matrices $W \in \mathbb{R}^{d_y \times d_x}$ of rank r which at $W \in \mathcal{M}_r$ has tangent space

$$T_W(\mathcal{M}_r) = \left\{ WA + BW : A \in \mathbb{R}^{d_x \times d_x}, B \in \mathbb{R}^{d_y \times d_y} \right\}.$$

It is shown in [2] that the operator

$$\mathcal{A}_W(Z) = \sum_{j=1}^{N} (WW^T)^{\frac{N-j}{N}} \cdot Z \cdot (W^T W)^{\frac{j-1}{N}}$$

is self-adjoint from $T_W(\mathcal{M}_r)$ into $T_W(\mathcal{M}_r)$, hence, invertible. Denoting its restriction to $T_W(\mathcal{M}_r)$ by $\overline{\mathcal{A}}_W$ the expression

$$g_W(Z_1, Z_2) = \langle \overline{\mathcal{A}}_W^{-1}(Z_1), Z_2 \rangle, \quad Z_1, Z_2 \in T_W(\mathcal{M}_r),$$

defines a Riemannian metric of class C^1. With the corresponding Riemannian gradient, the flow (3) becomes a Riemannian gradient flow [2].

Extending [3] it is shown in [2] that the gradient flow (2) always converges to a critical point of L^N. Moreover, for almost all initializations the flow converges to a global minimum of L^N – see [2] for the precise statement. These results have been generalized in [7] to the gradient descent (1) under an upper bound on the step sizes η_k.

As already mentioned above, in overparameterized settings, where the global minimizer is not unqiue, it has been observed that gradient descent on linear networks favors solutions of low rank, see e.g. [5, 6, 8]. In order to study this phenomenon, the dynamics of gradient flow and dynamics for minimizing the functional

$$K^N(W_N, \ldots, W_1) = \frac{1}{2} \| W_N \cdots W_1 - \widehat{W} \|_F^2,$$

where $\widehat{W}$ is a given symmetric matrix, has been analyzed in [4] for the special identical initialization $W_j(0) = \alpha \, \mathrm{Id}$, where $\alpha > 0$ is a suitable small constant. For $N \geq 2$ it turns out that the trajectory of the product $W(k) = W_N(k) \cdots W_1(k)$ first approaches the best rank one approximation of $\widehat{W}$, then the best rank two approximation and so on. Precise intervals are provided, in which one is close to a given rank k approximation.

Many open problems remain such as extending the previously mentioned result to random initializations and to a matrix sensing problem with many global minimizers. Of course, a challenging main goal is to develop a rigorous analysis for the general case of nonlinear neural networks.

References

[1] S. Arora, N. Cohen, N. Golowich, and W. Hu. *A convergence analysis of gradient descent for deep linear neural networks.* In International Conference on Learning Representations, 2019.

[2] B. Bah, H. Rauhut, U. Terstiege, and M. Westdickenberg. *Learning deep linear neural networks: Riemannian gra- dient flows and convergence to global minimizers.* Information and Inference: A Journal of the IMA, to appear. DOI:10.1093/imaiai/iaaa039.

[3] Y. Chitour, Z. Liao, and R. Couillet. *A geometric approach of gradient descent algorithms in neural networks.* Preprint arXiv:1811.03568, 2018.

[4] H.H. Chou, C. Gieshoff, J. Maly, H. Rauhut. *Gradient Descent for Deep Matrix Factorization: Dynamics and implicit bias towards low rank* Preprint arXiv:2011.13772, 2020.

[5] D. Gissin, S. Shalev-Shwartz, and A. Daniely. *The implicit bias of depth: How incremental learning drives generalization.* In International Conference on Learning Representations (ICLR), 2020.

[6] S. Gunasekar, B. E. Woodworth, S. Bhojanapalli, B. Neyshabur, and N. Srebro. *Implicit regularizationin matrix factorization.* In Advances in Neural Information Processing Systems, pages 6151–6159, 2017.

[7] G.M. Nguegnang, H. Rauhut, and U. Terstiege. *Convergence of gradient descent for learning linear neural networks.* Preprint arXiv:2108.02040, 2021.

[8] N. Razin and N. Cohen. *Implicit regularization in deep learning may not be explainable by norms.* InAdvances in Neural Information Processing Systems, 2020.

[9] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. *Understanding deep learning requires rethinking generalization.* In International Conference on Learning Representations, 2017.

Space-time Methods for Parabolic Evolution Equations

Rob Stevenson

(joint work with W. Dahmen, G. Gantner, R. van Venetië, J. Westerdiep)

Parabolic evolution equations in a simultaneous space-time variational formulation. Let V, H be separable Hilbert spaces of functions on some "spatial domain" such that $V \hookrightarrow H$ with dense embedding. Identifying H with its dual, we obtain the Gelfand triple $V \hookrightarrow H \simeq H' \hookrightarrow V'$. We use $\langle \cdot, \cdot \rangle$ to denote both the scalar product on $H \times H$ as well as its unique extension to the duality pairing on $V' \times V$ or $V \times V'$, and denote the norm on H by $\| \cdot \|$.

For a.e. $t \in I := (0,T)$, let $a(t;\cdot,\cdot)$ denote a bilinear form on $V \times V$ such that for any $\eta, \zeta \in V$, $t \mapsto a(t;\eta,\zeta)$ is measurable on I, and for a.e. $t \in I$, $a(t;\cdot,\cdot)$ is *bounded* and *coercive.*

With $A(t) \in \mathcal{L}\mathrm{is}(V, V')$ defined by $(A(t)\eta)(\zeta) := a(t;\eta,\zeta)$, given a forcing function g and an initial value u_0, we are interested in solving the *parabolic initial value problem* of finding u such that

$$\begin{cases} \frac{du}{dt}(t) + A(t)u(t) &= g(t) \quad (t \in I), \\ u(0) &= u_0. \end{cases}$$

In a simultaneous space-time variational formulation, the parabolic PDE reads as finding u from a suitable space of functions X of time and space that satisfies

$u(0) = u_0$ and

$$(Bw)(v) := \int_I \langle \tfrac{dw}{dt}(t), v(t)\rangle + a(t; w(t), v(t))dt = \int_I \langle g(t), v(t)\rangle \, dt =: g(v)$$

for all v from another suitable space of functions Y of time and space.

Theorem 1 (e.g. [2, Ch.XVIII, §3] or [7, Ch. IV, §26]). *With $X := L_2(I; V) \cap H^1(I; V')$, $Y := L_2(I; V)$, it holds that*

$$(B, \gamma_0) \in \mathcal{L}\mathrm{is}(X, Y' \times H),$$

with $\gamma_t \colon u \mapsto u(t, \cdot)$ denoting the trace map.

We define $A, A_s \in \mathcal{L}\mathrm{is}(Y, Y')$, $A_a \in \mathcal{L}(Y, Y')$ by

$$(Aw)(v) := \int_I a(t; w(t), v(t))\, dt, \quad A_s := \tfrac{1}{2}(A + A'), \quad A_a := \tfrac{1}{2}(A - A'),$$

and equip Y with 'energy'-scalar product $\langle \cdot, \cdot \rangle_Y := (A_s \cdot)(\cdot)$, and norm

$$\|v\|_Y := \sqrt{(A_s v)(v)}.$$

being equivalent to the standard norm on Y. We equip Y' with the resulting dual norm. We equip X with norm

$$\| \cdot \|_X := \sqrt{\| \cdot \|_Y^2 + \|\partial_t \cdot \|_{Y'}^2 + \|\gamma_T \cdot \|^2},$$

being, thanks to $X \hookrightarrow C(\overline{I}; H)$, equivalent to the standard norm on X.

Minimal residual discretization. Let $(X^\delta, Y^\delta)_{\delta \in \Delta}$ be a family of closed subspaces of X and Y, respectively. For $\delta \in \Delta$, let E_X^δ and E_Y^δ denote the trivial embeddings $X^\delta \to X$ and $Y^\delta \to Y$. We assume that

$$X^\delta \subseteq Y^\delta \quad (\delta \in \Delta), \tag{1}$$

$$\gamma_\Delta^{\partial_t} := \inf_{\delta \in \Delta} \inf_{\{w \in X^\delta \colon \partial_t E_X^\delta w \neq 0\}} \frac{\|{E_Y^\delta}' \partial_t E_X^\delta w\|_{Y^{\delta\prime}}}{\|\partial_t E_X^\delta w\|_{Y'}} > 0. \tag{2}$$

Our Minimal Residual approximation $u^\delta \in X^\delta$ of the solution $u \in X$ of $(B, \gamma_0)u = (g, u_0)$ is defined as

$$u^\delta := \operatorname*{argmin}_{w \in X^\delta} \|{E_Y^\delta}'(B E_X^\delta w - g)\|_{Y^{\delta\prime}}^2 + \|\gamma_0 E_X^\delta w - u_0\|^2, \tag{3}$$

The numerical approximation (3) was proposed in [1], and further investigated in [5]. So far the analysis of the MR method was restricted to the case that $A_a = 0$.

Theorem 2 ([6]). *Under conditions* (1) *and* (2), *and with $\alpha := \|A_a\|_{\mathcal{L}(Y,Y')} = \rho(A_s^{-1} A_a A_s^{-1} A_a)^{\frac{1}{2}}$, the solution $u^\delta \in X^\delta$ of* (3) *exists uniquely, and satisfies*

$$\|u - u^\delta\|_X \leq \sqrt{\frac{\left(1 + \frac{1}{2}\left(\alpha^2 + \alpha\sqrt{\alpha^2+4}\right)\right)}{\frac{1}{2}\left((\gamma_\Delta^{\partial_t})^2 + \alpha^2 + 1 - \sqrt{((\gamma_\Delta^{\partial_t})^2 + \alpha^2 + 1)^2 - 4(\gamma_\Delta^{\partial_t})^2}\right)}} \inf_{w \in X^\delta} \|u - w\|_X.$$

Wavelets-in-time, finite elements-in-space. In [1], (1) and (2) were verified for $(X^\delta, Y^\delta)_{\delta\in\Delta}$ being families of pairs 'full' and 'sparse' tensor products of finite element spaces in time and space. These families however do not accommodate local refinements simultaneous in space and time.

To allow for such refinements, in [4] we consider the following setting: Let $\Sigma = \{\sigma_\lambda \colon \lambda \in \vee_\Sigma\}$ be a wavelet Riesz basis for $L_2(I)$, such that $\{2^{-|\lambda|}\sigma_\lambda \colon \lambda \in \vee_\Sigma\}$ is a Riesz basis Riesz for $H^1(I)$. To $\lambda \in \vee_\Sigma$ with $|\lambda| > 0$, we associate $\tilde{\lambda} \in \vee_\Sigma$ with $|\tilde{\lambda}| = |\lambda| - 1$ and $\operatorname{supp} \sigma_\lambda \subseteq \operatorname{supp} \sigma_{\tilde{\lambda}}$. We denote this parent-child relation by $\tilde{\lambda} \lhd_\Sigma \lambda$. Let $\Psi = \{\psi_\lambda \colon \lambda \in \vee_\Psi\}$ orthonormal basis for $L_2(I)$ such that

$$\operatorname{span}\{\sigma_\lambda \colon |\lambda| \le \ell\} \cup \operatorname{span}\{\sigma'_\lambda \colon |\lambda| \le \ell\} \subseteq \operatorname{span}\{\psi_\mu \colon |\mu| \le \ell\}.$$

Let $\mathcal{O}$ be a collection of of finite element spaces in V, closed under taking (finite) unions, with

$$\inf_{W\in\mathcal{O}} \inf_{0\neq w\in W} \sup_{0\neq v\in W} \frac{w(v)}{\|w\|_{V'}\|v\|_V} > 0$$

which condition is equivalent to uniform boundedness in V of the H-orthogonal projector onto $W \in \mathcal{O}$. For the model case of $H = L_2(\Omega)$, $V = H^1_0(\Omega)$ and Ω a polygon, this condition is known to be satisfied for the family of continuous piecewise linears, zero on $\partial\Omega$, w.r.t. all conforming triangulations that can be generated by newest vertex bisection starting from a conforming initial triangulation of Ω with an assignment of the newest vertices that satisfies a matching condition.

Proposition 3 ([4]). *Let* $X^\delta = \sum_{\lambda\in\vee_\Sigma} \sigma_\lambda \otimes W^\delta_\lambda$ *such that* $\tilde{\lambda} \lhd_\Sigma \lambda$ *implies* $W^\delta_{\tilde{\lambda}} \supseteq W^\delta_\lambda$. *Take* $Y^\delta = \sum_{\mu\in\vee_\Psi} \psi_\mu \otimes \bar{W}^\delta_\mu$ *with* $\bar{W}^\delta_\mu = \sum_{\{\lambda\in\vee_\Sigma \colon |\lambda|=|\mu|, |\operatorname{supp}\psi_\mu \cap \operatorname{supp}\sigma_\lambda|>0\}} W^\delta_\lambda$. *Then* (1) *and* (2) *are satisfied.*

Furthermore for this family $(X^\delta, Y^\delta)_{\delta\in\Delta}$, in [4] we show how to apply the system matrix resulting from the minimal residual discretization in linear complexity, build optimal preconditioners at X- and Y-side from multigrid-in-space, and design a r-linearly convergent adaptive algorithm under a saturation assumption. The computational cost and memory consumption of this algorithm is proportional to the dimension of the trial space X^δ. Numerical results with piecewise linear wavelets-in-time and piecewise linear finite element-in-space show for non-smooth solutions of the heat equation a considerable speed-up compared to non-adaptive full and sparse grid discretizations.

Data assimilation for parabolic problems. Let $H = L_2(\Omega)$ and $V = H^1_0(\Omega)$. Given $\emptyset \neq \omega \subset \Omega$, $g \in Y'$, $f \in L_2(I \times \omega)$, with $\Gamma_\omega w := w|_{I\times\omega}$ consider the data-assimilation problem of finding $u \in X$ with $(B, \Gamma_\omega)u = (g, f)$. It holds that $(B, \Gamma_\omega) \in \mathcal{L}(X, Y' \times L_2(I \times \omega))$, but (B, Γ_ω) is not surjective, so data can be inconsistent, and although (B, Γ_ω) is injective, it holds that $\|w\|_X \not\lesssim \|Bw\|_{Y'} + \|\Gamma_\omega w\|_{L_2(I\times\omega)}$. What does hold is that with $X_\eta := L_2((\eta, T); V) \cap H^1((\eta, T); V')$, for any fixed $\eta > 0$, $\|w\|_{X_\eta} \lesssim \|Bw\|_{Y'} + \|\Gamma_\omega w\|_{L_2(I\times\omega)}$ (*Carleman estimate*).

Theorem 4 ([3]). *Assuming* $\inf_{\delta\in\Delta}\inf_{\{w\in X^\delta\colon Bw\neq 0\}}\frac{\|E_Y^\delta{}' B E_X^\delta w\|_{Y^\delta{}'}}{\|BE_X^\delta w\|_{Y'}} > 0$*, then for* $u \in X$ *and* $\varepsilon \geq 0$*,*

$$u^\delta := \operatorname*{argmin}_{w\in X^\delta} \|E_Y^\delta{}' BE_X^\delta w - g\|_{Y^\delta{}'}^2 + \|\Gamma_\omega E_X^\delta w - f\|_{L_2(I\times\omega)}^2 + \varepsilon^2 \|\gamma_0 E_X^\delta w\|_{L_2(\Omega)}^2,$$

$$\|u - u^\delta\|_{X_\eta} \lesssim \|Bu - g\|_{Y'} + \|\Gamma_\omega u - f\|_{L_2(I\times\omega)} + \min_{w\in X^\delta} \|u - w\|_X + \varepsilon\|\gamma_0 u\|_{L_2(\Omega)}.$$

References

[1] R. Andreev. *Stability of sparse space-time finite element discretizations of linear parabolic evolution equations.* IMA J. Numer. Anal., 33(1):242–260, 2013. doi:10.1093/imanum/drs014.

[2] R. Dautray and J.-L. Lions. *Mathematical analysis and numerical methods for science and technology. Vol. 5.* Springer-Verlag, Berlin, 1992. Evolution problems I. doi:10.1007/978-3-642-58090-1.

[3] W. Dahmen, R.P. Stevenson, and J. Westerdiep. *Accuracy controlled data assimilation for parabolic problems,* 2021. arXiv:2105.05836.

[4] R.P. Stevenson, R. van Venetië, and J. Westerdiep. *A wavelet-in-time, finite element-in-space adaptive method for parabolic evolution equations*, 2021. arXiv:2101.03956.

[5] R.P. Stevenson and J. Westerdiep. *Stability of Galerkin discretizations of a mixed space-time variational formulation of parabolic evolution equations.* IMA J. Numer. Anal., 2020. URL: `https://doi:10.1093/imanum/drz069`.

[6] R.P. Stevenson and J. Westerdiep. *Minimal residual space-time discretizations of parabolic equations: Asymmetric spatial operators,* 2021. arXiv:2106.01090.

[7] J. Wloka. *Partielle Differentialgleichungen.* B. G. Teubner, Stuttgart, 1982. Sobolevräume und Randwertaufgaben.

Shallow Thoughts on Deep Learning

Ronald A. DeVore

Deep Learning (DL) is very much in the news these days. It has had a myriad of empirical successes. However, from the viewpoint of a mathematician, the reasons behind these successes are not clear and a rigorous mathematical theory to certify the performance of deep learning algorithms is still lacking. Indeed, many of the components of deep learning algorithms are counter intuitive. One component usually lacking in DL algorithms is the appearance of a model class assumption on the function to be learned. In the absence of such knowledge we argue that no quantitative performance bounds can be derived.

This talk gives a mathematical view of deep learning algorithms and examines its main components. We begin by showing that the canonical learning problem has a theoretical optimal solution described by a certain Chebyshev ball described by the data and the information provided about the target function. However, it is highly nontrivial to turn this theoretical solution into a numerical algorithm.

We next discuss possible numerical algorithms that can be employed when a model class assumption is present. This includes least squares, penalized leased square, and LASSO algorithms. We then turn to the justification in the DL community of the absence of such model class assumptions. Here, the rationale is that the implementations of numerical optimization procedures such as gradient descent and stochastic gradient descent implicitly impose a model class assumption.

An important ingredient in deep learning is the employment of neural networks as the numerical backbone of the algorithm. We discuss the possible advantages of using neural networks as the approximation tool. Neural networks have expanded approximation properties. In addition to approximating classical model classes of functions well, they have the capacity to efficiently approximate classes of functions described by self-similarity and dynamical systems. A point to emphasize however is that neural network approximation can be very unstable and therefore implementation must be done with care.

As a final topic of this talk we discuss generalization error and its efficacy in guaranteeing performance. While, in some settings, generalization error can provide certifiable guarantees in expectation when using random sampling, these guarantees fall short of what can be obtained when employing a priori model class assumptions. We point out the large gap in our understanding on this topic.

On the Universality of Gradient Descent for Neural Network Training and Global Minimizers in non-convex Compressed Sensing by meta-learning

Gerrit Welper

Successful neural network training commonly requires a judicious choice of hyperparameters and network architecture. Indeed many common neural network components such as ReLU activations, skip connections or LSTM units have been introduced to improve training performance. Likewise, we discuss training results, which allow the extra flexibility to redesign the network for a class of learning problems, unlike contemporary training theory, which fixes the network architecture and provides convergence guarantees in over-parametrized regimes. This approach yields the following universality result [2]: Assume that there exists a training algorithm that produces satisfactory weights for a neural network on a range of learning problems. Then there exists an extension of the network so that gradient descent training yields the same forward model, as a function. I.e., after training with the respective methods, for every input both networks produce the same output, while their internal computations may differ.

Similar to universal approximation theorems, the extended network is handcrafted and therefore impractical. Nonetheless, more practical analogues are provided by neural architecture search and transfer-, multitask- and meta-learning. In these problems, architectures or a subset of neural network weights are pretrained on a class of problems and then fine tuned or specialized to instances of the problem class. The extended network in the universality result serves as an idealized candidate of the pre-training outcome, and demonstrates the potential of the meta-learning approach.

While proving that practical pre-training can achieve similar results than the idealized extended network is expected to be extremely difficult, we provide some proof of principle for simplified non-convex model problems from compressed sensing [1, 3]. To this end, we consider a class of non-convex ℓ_0 optimization problems, which contains some instances that are "easy" and can be solved by the usual ℓ_1 relaxation and some instances that are "hard" and cannot be solved by known efficient algorithms. In addition, we assume that the class instances are related by some unknown model. Using only samples form the problem class, consisting of measurement matrices A and right hand sides b for sparse linear systems $Ax = b$, but no solutions x, in a pre-learning phase, we uncover the relations of the class problems. This step is comparable to finding the extended network in the universaility result above. After pre-training, we can use the class' structure to solve every problem in class (with high probabilty), including the "hard" ones for which we initially had no efficient algorithm. This part is comparable to the gradient descent training of the extended network on a class instance. In summary, given samples from a mixture of related "easy" and "hard" problems, we can learn to solve all problems in class.

References

[1] G. Welper, *A Relaxation Argument for Optimization in Neural Networks and Non-Convex Compressed Sensing*, `https://arxiv.org/abs/2002.00516`, 2020.

[2] G. Welper, *Universality of Gradient Descent Neural Network Training*, `https://arxiv.org/abs/2007.13664`, 2020.

[3] G. Welper, *Non-Convex Compressed Sensing with Training Data*, `https://arxiv.org/abs/2101.08310`, 2021.

Sampling Rates for ℓ^1-Synthesis

Claire Boyer

(joint work with Jonas Kahn, Maximilian März, Pierre Weiss)

This work investigates the problem of the recovery of a signal $x_0 \in \mathbb{R}^n$ from m undersampled ($m \ll n$) noisy sub-Gaussian measurements:

$$y = Ax_0 + e$$

where $y \in \mathbb{R}^m$ is the observation vector, $A \in \mathbb{R}^{m\times n}$ is a sugGaussian matrix, and $e \in \mathbb{R}^m$ is a noise vector. While a lot of attention has been given to the analysis formulation in the literature, we study the synthesis-based sparsity model such that the signal x_0 is assumed to be synthesized by some atoms of a dictionary $D \in \mathbb{R}^{n\times d}$, i.e.

$$x_0 = Dz_0$$

with $z_0 \in \mathbb{R}^d$ the coefficient vector. Solving the l1-synthesis basis pursuit allows to simultaneously estimate a coefficient representation as follows:

$$\hat{Z} := \underset{z\in\mathbb{R}^d}{argmin} \|z\|_1 \quad \text{such that} \quad \begin{cases} y = ADz & \text{in the noiseless setting,} \\ \|y - ADz\|_2 \leq \eta & \text{in the noisy one.} \end{cases}$$

as well as the sought-for signal by applying the dictionary D to the latter set of solutions:

$$\hat{X} := D\hat{Z}.$$

While these programs are pretty well understood in the case where D is orthogonal, choosing a redundant dictionary (which would be done in practice) is a different story. Indeed, due to linear dependencies within redundant dictionary atoms it might be impossible to identify a specific representation vector, although the actual signal is still successfully recovered. We study both estimation problems from a non-uniform, signal-dependent perspective. By utilizing results from linear inverse problems and convex geometry, we identify the sampling rate describing the phase transition of both formulations: the phase transition is determined by the Gaussian width, denoted by ω, of a linearly-transformed polyhedral cone,

$$m \geq c \cdot \omega^2 \left(D\mathcal{D}(\|\cdot\|_1, z_{\ell^1}) \right),$$

where $\mathcal{D}(\|\cdot\|_1, z_{\ell^1})$ is the descent cone of the ℓ^1-norm at a minimal-ℓ^1-representer $z_{\ell^1} \in argmin_z\{\|z\|_1 : x_0 = Dz\}$ of the signal x_0. We provide a "tight" upper bound on this Gaussian width, performing a geometric analysis of the thinness of high-dimensional polyhedral cones with not exponentially many generators. We believe that such an argument might be of general interest beyond its application to the synthesis formulation of compressed sensing. Again, $\omega^2 \left(D\mathcal{D}(\|\cdot\|_1, z_{\ell^1}) \right)$ is related to the sparsity of a minimal ℓ^1-representation and a further geometrical parameter (referred to as the *circumangle*) that measures the narrowness of the associated cone. An important aspect of this bound is that its computation boils down to a convex optimization problem, which is numerically tractable. In addition, it can be evaluated analytically in some situations of interest. This enables us to demonstrate its usefulness in several examples and to identify non-trivial situations in which such a result is asymptotically near-optimal.

References

[1] M. März, C. Boyer, J. Kahn, P. Weiss, *Sampling rates for ℓ^1-synthesis*, arXiv:2004.07175, 2020.

A near-optimal Adaptive Stochastic Galerkin Method based on Multilevel Expansions of Random Fields

Markus Bachmayr
(joint work with Igor Voulis)

We consider elliptic PDEs depending on infinitely many parameters entering into a parametrized series expansion of the diffusion coefficient. Problems of this type arise in particular in the deterministic approximation of PDEs with random diffusion coefficients. The variational formulation of the considered affinely parameterized elliptic PDEs on a domain $D \subset \mathbb{R}^d$ reads: $u(y) \in V := H_0^1(D)$ such that

$$\int_D \Big(\theta_0 + \sum_{j=1}^{\infty} y_j \theta_j\Big) \nabla u(y) \cdot \nabla v \,\mathrm{d}x = f(v), \quad v \in V, \quad \text{for } y \in Y := [-1,1]^{\mathbb{N}},$$

where $f \in V'$, $\theta_j \in L_\infty(D)$ for $j \in \mathbb{N}$ and $\theta_0 \in L_\infty(D)$ are such that the problem is elliptic uniformly in y. For simplicity, we assume the scalar coefficients y_j to be uniformly distributed in $[-1,1]$, and accordingly study approximation of u in the Bochner space $L_2(Y, V, \sigma)$ with σ the uniform measure on Y, but other product measures can be treated in the same manner.

The focus of this work is on adaptive algorithms for computing sparse Legendre approximations of u. We use the orthonormal basis $\{L_\nu\}_{\nu\in\mathcal{F}}$ of $L_2(Y,\sigma)$, where $L_\nu(y) = \prod_{j=1}^\infty L_{\nu_j}(y_j)$ are the product Legendre polynomials and $\mathcal{F}$ denotes the multi-indices in $\mathbb{N}_0^{\mathbb{N}}$ of finite support. By truncating the expansion of u with respect to this orthonormal basis to some $F \subset \mathcal{F}$ of finite support, we obtain a *semi-discrete* approximation

$$u(y) \approx u_F(y) = \sum_{\nu \in F} u_\nu L_\nu(y), \tag{1}$$

where each Legendre coefficient u_ν is a function in V. In order to obtain a fully discrete, numerically computable approximation, each u_ν in turn needs to be replaced by an approximation, which we assume (as in typical approximations by finite elements or wavelets) to be chosen from some finite-dimensional subspace $V_\nu \subset V$. The total number of degrees of freedom in such a fully discrete approximation u_N is then $N = \sum_{\nu\in F} \dim(V_\nu)$.

When the functions θ_j have multilevel structure with a scale parameter $\ell(j)$ such that $\operatorname{diam}\operatorname{supp}\theta_j \lesssim 2^{-\ell(j)}$ for each $j \in \mathbb{N}$ – for instance, when they correspond to a suitably rescaled wavelet-type basis – one obtains improved convergence results for such Legendre expansions. As shown in [4], if for all scales $\hat\ell$,

$$\Big\| \sum_{\ell(j)=\hat\ell} |\theta_j| \Big\|_{L_\infty} \lesssim 2^{-\alpha\hat\ell}, \qquad \#\{j \colon \ell(j) = \hat\ell\} \lesssim 2^{d\hat\ell}, \tag{2}$$

then best approximations of the form (1) converge as $\|u - u_F\|_{L_2(Y,V,\sigma)} \lesssim (\#F)^{-s}$ for any $s < \alpha/d$. Estimates with natural additional conditions on derivatives of the θ_j indicate the potential advantages of using *independent* adaptive spatial discretizations for each Legendre coefficient: as shown in [2], for $d \geq 2$ and $\alpha \in (0,1]$, with standard adaptive spatial approximations, there exist fully discrete approximations u_N such that $\|u - u_N\|_{L_2(Y,V,\sigma)} \lesssim N^{-s}$ for any $s < \alpha/d$ (in the exceptional case $d = 1$, this result was shown only for any $s < \frac{2}{3}\alpha$). For $d \geq 2$, this result means that compared to semidiscrete approximations, the additional spatial discretization comes at no additional expense in the convergence rate.

In this work, we address the question whether an adaptive algorithm can be constructed that finds such fully discrete approximations at optimal computational cost, that is, using $\mathcal{O}(N)$ operations. It was previously shown in [3] that with adaptive wavelet schemes for the spatial discretization, one can obtain rates that are close to optimal under additional regularity requirements on the basis functions that are used. These requirements, however, are too strong to be practical. In our new approach, combining adaptive operator compression for the parametric expansion with spline wavelet tree approximation as in [6] for the spatial coefficients, we obtain a method converging at optimal rates and requiring $\mathcal{O}(N\log N)$ operations, thus achieving near-optimal complexity, under natural assumptions.

Each space V_ν is chosen as the span of a finite subset $S_\nu \subset \mathcal{S}$ of a fixed spatial spline wavelet basis $\{\psi_\lambda\}_{\lambda\in\mathcal{S}}$, with the additional constraint that each S_ν has a certain tree structure. The original problem for u can be recast as approximating the corresponding coefficient sequence $\mathbf{u} = (\mathbf{u}_{\nu,\lambda})_{\nu\in\mathcal{F},\lambda\in\mathcal{S}}$ of u in the product basis $\{\psi_\lambda \otimes L_\nu\}_{\nu\in\mathcal{F},\lambda\in\mathcal{S}}$ in $\ell_2(\mathcal{F}\times\mathcal{S})$. Fully discrete approximations with the additional tree restriction converge at rate $s > 0$ if

$$\|\mathbf{u}\|_{\mathrm{t},s} = \sup_{N\in\mathbb{N}_0} (N+1)^s \min\bigl\{ \|\mathbf{u}-\mathbf{v}\|_{\ell_2} \colon \operatorname{supp}\mathbf{v} \subseteq \{(\nu,\lambda)\colon \nu\in F, \lambda\in S_\nu\}, \\ S_\nu \subset \mathcal{S} \text{ tree for } \nu\in F,\ \sum_\nu \#S_\nu \le N \bigr\} < \infty.$$

We first show that the same rates can be achieved by this more constrained type of approximation as in the original results of [2].

The adaptive scheme follows the basic strategy of successively refined Galerkin discretizations of [5, 6]. The core element of the method is a new technique for approximating full spatial-parametric residuals that makes crucial use both of the multilevel structure (2) of the parameterization of the random coefficient and of the piecewise polynomial structure of basis functions. The selection of the most relevant degrees of freedom from the residual approximations is then done by an adaptation of the quasi-optimal tree coarsening analyzed in [1]. Our main result on the new adaptive method is the following.

Theorem. *let $f \in L_2(D)$, let $\theta_j \in C^{0,1}$ for $j\in\mathbb{N}$ with multiscale structure (2), let $\psi_\lambda \in H^2$ for $\lambda\in\mathcal{S}$, and let each of these functions be piecewise polynomial with respect to a joint sequence of tesselations of D. Let $0 < s < \frac{\alpha}{d}$ and $\|\mathbf{u}\|_{\mathrm{t},s} < \infty$. Then for each $\varepsilon > 0$, the adaptive scheme with appropriately chosen parameters finds an approximation $\mathbf{u}^k$ for some $k\in\mathbb{N}$ with $\|\mathbf{u}-\mathbf{u}^k\|_{\ell_2} \le \varepsilon$, such that:*

(i) There exists $C > 0$ independent of ε, but depending on s, such that

$$\#\operatorname{supp}\mathbf{u}^k \le C\,\varepsilon^{-\frac{1}{s}}\,\|\mathbf{u}\|_{\mathrm{t},s}^{\frac{1}{s}}.$$

(ii) The computation can be realized to use a number of operations of order

$$(\#\operatorname{ops}\mathbf{u}^k) \lesssim 1 + \varepsilon^{-\frac{1}{s}}\|\mathbf{u}\|_{\mathrm{t},s}^{\frac{1}{s}}\bigl(1 + |\log\varepsilon| + \log\|\mathbf{u}\|_{\mathrm{t},s}\bigr)\,.$$

Preliminary numerical tests confirm the expected convergence rates, and in particular they indicate that not only the rates for $d\ge 2$, but also the rate of $\frac{2}{3}\alpha$ obtained for $d=1$ in [2] are sharp and that the for all d, the respective convergence rates results extend to $\alpha > 1$.

References

[1] P. Binev, *Tree approximation for hp-adaptivity*, SIAM J. Numer. Anal. **56** (2018), 3346–3357.

[2] M. Bachmayr, A. Cohen, D. Dũng, and C. Schwab, *Fully discrete approximation of parametric and stochatic elliptic PDEs*, SIAM J. Numer. Anal. **55** (2017), 2151–2186.

[3] M. Bachmayr, A. Cohen, and W. Dahmen, *Parametric PDEs: Sparse or low-rank approximations?*, IMA J. Numer. Anal. **38** (2018), 1661–1708.

[4] M. Bachmayr, A. Cohen, and G. Migliorati, *Sparse polynomial approximation of parametric elliptic PDEs. Part I: affine coefficients*, ESAIM Math. Model. Numer. Anal. **51** (2017), 321–339.

[5] T. Gantumur, H. Harbrecht, and R. Stevenson, *An optimal adaptive wavelet method without coarsening of the iterands*, Math. Comp. **76** (2007), 615–629.

[6] R. Stevenson, *Adaptive wavelet methods for linear and nonlinear least-squares problems*, Found. Comput. Math. **14** (2014), 237–283.

Multilevel Approximation of Gaussian Random Fields

Helmut Harbrecht

(joint work with Lukas Herrmann, Kristin Kirchner, Michael Multerer, and Christoph Schwab)

1. Introduction

Centered Gaussian random fields are uniquely determined by their covariance operator $\mathcal{C}$ or their precision operator $\mathcal{P} = \mathcal{C}^{-1}$. Throughout this work, we consider Gaussian random fields $\mathcal{Z}$ which are generated by a linear colouring (elliptic pseudo-differential) operator $\mathcal{A} \in OPS^r_{1,0}(\mathcal{M})$ of order $r > n/2$ via the white noise driven stochastic (pseudo-) differential equation

$$\mathcal{A}\mathcal{Z} = \mathcal{W} \text{ on } \mathcal{M}.$$

Here, $\mathcal{M}$ is a smooth domain in $\mathbb{R}^n$ or a smooth manifold in $\mathbb{R}^{n+1}$ and $\mathcal{W}$ denotes the white noise in $L^2(\mathcal{M})$. If $\mathcal{A} : H^{r/2}(\mathcal{M}) \to H^{-r/2}(\mathcal{M})$ is self-adjoint and positive

$$\langle \mathcal{A}w, w\rangle \gtrsim \|w\|^2_{r/2} \text{ for all } w \in H^{r/2}(\mathcal{M}) \setminus \{0\},$$

then it holds

$$\mathcal{C} = \mathcal{A}^{-2} \in OPS^{-2r}_{1,0} \text{ and } \mathcal{P} = \mathcal{A}^2 \in OPS^{2r}_{1,0}.$$

Note that this setup includes the Matérn class of covariance operators.

Several methodologies in uncertainty quantification and data assimilation require the storage of the covariance matrix $\mathbf{C}$ or the precision matrix $\mathbf{P} = \mathbf{C}^{-1}$ corresponding to an underlying statistical model as well as computations involving these matrices. Explicit examples include simulations, predictions and Bayesian or likelihood-based inference in spatial statistics. Here, one of the main computational challenges is to handle large datasets, as the covariance and precision matrices $\mathbf{C}$, $\mathbf{P}$ are, in general, densely populated and, for this reason, the computational cost for predictions or inference is cubic in the number of observations.

2. Wavelet matrix compression

The Schwartz kernel $\kappa_{\mathcal{B}}$ of a pseudo-differential operator $\mathcal{B} \in OPS^r_{1,0}$ is asymptotically smooth, meaning that

$$\left|\partial_{\mathbf{x}}^{\boldsymbol{\alpha}}\partial_{\mathbf{y}}^{\boldsymbol{\beta}}\kappa_{\mathcal{B}}(\mathbf{x},\mathbf{y})\right| \leq c_{\boldsymbol{\alpha},\boldsymbol{\beta}}\|\mathbf{x}-\mathbf{y}\|^{-(n+r+|\boldsymbol{\alpha}|+|\boldsymbol{\beta}|)}.$$

Such an estimate is the key to compress the discretized version of the covariance or precision operator by means of wavelet matrix compression. In accordance with [2], by using appropriate wavelet bases, we can represent such operators with linear cost while preserving discretization error accuracy. Indeed, it is proven in [3] that the covariance and precision operators, respectively, may be identified with bi-infinite matrices and finite sections may be diagonally preconditioned rendering the condition number independent of the dimension N of this section. As an illustration, we consider the Matérn covariance kernel $k_{1/2}(\mathbf{x},\mathbf{y}) = \exp(-\|\mathbf{x}-\mathbf{y}\|)$ on the boundary curve of a smooth two-dimensional domain. The results for (periodic) biorthogonal wavelets $\psi_{d,\widetilde{d}}$ from [1] of order d and with $\widetilde{d}$ vanishing moments are found in Table 1, where we clearly see bounded condition numbers and asymptotically optimal compression rates.

$k_{1/2}$							
N	single-scale	nnz	$\psi^{(2,4)}$	nnz	$\psi^{(2,6)}$	nnz	$\psi^{(2,8)}$
32	$2.6 \cdot 10^3$	100	$2.4 \cdot 10^2$	100	$1.8 \cdot 10^2$	100	$6.6 \cdot 10^2$
64	$1.1 \cdot 10^4$	80	$2.7 \cdot 10^2$	88	$1.9 \cdot 10^2$	98	$6.7 \cdot 10^2$
128	$4.5 \cdot 10^4$	60	$3.1 \cdot 10^2$	65	$1.9 \cdot 10^2$	71	$6.8 \cdot 10^2$
256	$1.9 \cdot 10^5$	40	$3.4 \cdot 10^2$	42	$1.9 \cdot 10^2$	48	$6.8 \cdot 10^2$
512	$7.6 \cdot 10^5$	25	$3.7 \cdot 10^2$	26	$1.9 \cdot 10^2$	30	$6.8 \cdot 10^2$
1024	$3.1 \cdot 10^6$	16	$3.9 \cdot 10^2$	16	$1.9 \cdot 10^2$	18	$6.8 \cdot 10^2$
2048	$1.2 \cdot 10^7$	9.4	$4.0 \cdot 10^2$	9.0	$1.9 \cdot 10^2$	10	$6.8 \cdot 10^2$
4096	$5.0 \cdot 10^7$	5.0	$4.2 \cdot 10^2$	5.0	$1.9 \cdot 10^2$	5.7	$6.8 \cdot 10^2$

Table 1. Condition numbers and compression rates in case of the Matérn covariance kernel $k_{1/2}$. The compression rates validate the asymptotically linear behaviour. The condition numbers stay bounded for $\psi^{(2,6)}$ and $\psi^{(2,8)}$, whereas for $\psi^{(2,4)}$ a slight increase is observed.

Wavelet matrix compression of covariance and precision operators results in several powerful algorithms like fast sampling, multilevel Monte Carlo oracles for covariance estimation, and kriging. In particular, it has been observed in [4] that matrices in wavelet coordinates can directly be inverted by means of nested dissection. For example, the sparsity pattern of a Gaussian random field in case of $\mathcal{M} = \mathbb{S}^2$ can be found in the left plot of Figure 1. When applying nested dissection, one obtains the reordered sparsity pattern seen in the middle plot, for which the Cholesky factorization becomes computable and produces nearly no fill-in (right plot).

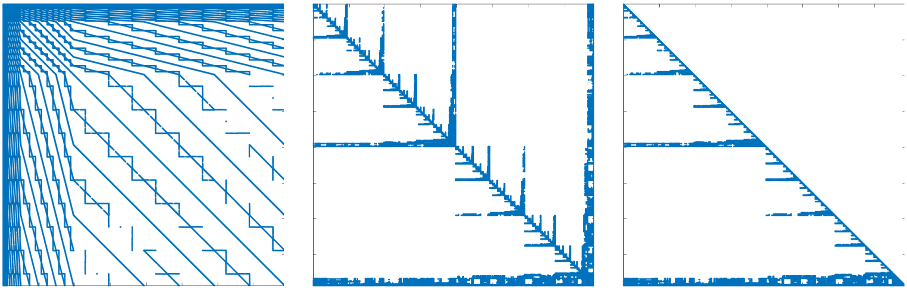

FIGURE 1. Sparsity pattern of the compressed covariance operator in wavelet coordinates (*left*), its nested dissection, “skyline” reordering (*middle*), and sparsity pattern of the exact Cholesky factor (*right*) of the compressed, reordered covariance matrix for $N = 393216$.

3. Samplets

Classically, wavelets are constructed by refinement relations and therefore require a sequence of nested approximation spaces which are copies of each other, except for a different scaling. This restricts the concept of wavelets to structured data. In order to generalize the concepts of wavelet matrix compression to discrete data, we introduce in [5] the concept of samplets which transfer the construction of Tausch-White wavelets [6] to the realm of data. This way, we obtain a multilevel basis which consists of localized and discrete signed measures. Inspired by the term wavelet, we call such signed measures *samplets*. Samplets can be constructed such that their associated measure integrals vanish for polynomial integrands.

When representing discrete data by samplets, then, due to the vanishing moments, there is a fast decay of the corresponding samplet coefficients with respect to the support size if the data are smooth. This directly enables data compression, detection of singularities and adaptivity. Applying samplets to represent covariance matrices and other kernel matrices, as they arise in kernel based learning or Gaussian process regression, we end up with quasi-sparse matrices. By thresholding small entries, these matrices are compressible to $\mathcal{O}(N \log N)$ relevant entries, where N is the number of data points. This feature allows for the use of fill-in reducing reorderings to obtain a sparse factorization of the compressed matrices. Besides the comprehensive introduction to samplets and their properties, extensive numerical studies in [5] demonstrate that samplets mark a considerable step in the direction of making large data sets accessible for analysis.

References

[1] A. Cohen, I. Daubechies, and J.-C. Feauveau, *Biorthogonal bases of compactly supported wavelets*, Comm. Pure Appl. Math. **45**(5) (1992) 485–560.

[2] W. Dahmen, H. Harbrecht, and R. Schneider, *Compression techniques for boundary integral equations. Asymptotically optimal complexity estimates*, SIAM J. Numer. Anal. **43**(6) (2006), 2251–2271.

[3] H. Harbrecht, L. Herrmann, K. Kirchner, and Ch. Schwab, *Multilevel approximation of Gaussian random fields: Covariance compression, estimation and spatial prediction*, SAM-Report 2021-09, ETH Zurich, Switzerland (2021).

[4] H. Harbrecht and M.D. Multerer, *A fast direct solver for nonlocal operators in wavelet coordinates*, J. Comput. Phys. **428** (2021), 110056.

[5] H. Harbrecht and M. Multerer, *Samplets: A new paradigm for data compression*, arXiv:2107.03337 (2021).

[6] J. Tausch and J. White, *Multiscale bases for the sparse representation of boundary integral operators on complex geometries*, SIAM J. Sci. Comput. **24** (2003), 1610–1629.

DeepOnets: A Machine Learning Framework in Infinite Dimensions

Siddhartha Mishra

A large number of problems involving PDEs are of the *many-query* type i.e., their solution requires multiple calls (queries) to an underlying numerical PDE solver, be it finite difference, finite element, finite volume or spectral. Such many-query problems arise in uncertainty quantification (UQ), deterministic and Bayesian inverse problems, optimal control and PDE constraint optimization. The biggest challenge in the robust and efficient solution of many-query problems is in the *computational cost.* As is well-known, traditional PDE solvers, particularly in three space dimensions, are very expensive. Hence, multiple calls to such PDE solvers lead to a prohibitively high computational cost, even on state of the art HPC platforms.

A possible approach for reducing computational cost is the use of *surrogate*, which approximate the data to solution map to high accuracy, but at a very small fraction of the cost of the underlying PDE solver. Recently, *machine learning* algorithms have been proposed in numerous papers as efficient surrogates for many-query problems in PDEs. In particular, deep neural networks are the most popular machine learning paradigms in this context.

This emerging paradigm rests on the availability of an efficient finite-dimensional parametrization of the input space, for instance, based on a Karhunen-Loeve expansion of the underlying measure. Then, the parameter to solution (or parameter to observable) map is approximated by a suitably trained deep neural networks. This approach has been very successful in many different contexts.

However, this approach rests on the assumption that a bespoke parametrization exists. This is far from the case in many contexts where PDEs arise. The underlying measure (a probability measure on a function space) may not be known to allow for the design of a good parametrization. At best, one can assume that it is possible to sample from this underlying measure. Moreover, deep neural network based surrogates are resolution dependent as they need to be trained on data generated by a traditional numerical method at a given resolution. Consequently, testing the trained network on a finer (or even coarser) resolution can lead to large errors.

Given this issues and realizing that most PDE learning tasks involve learning operators i.e., mappings between two infinite-dimensional Banach spaces, *operator learning* becomes an imperative for building surrogates for PDEs. Many different frameworks for learning operators from data have been proposed recently. In this talk, we described one class of such operator learning frameworks, namely deep operator networks or *DeepOnets*, for short.

We discussed results about DeepOnets that were presented in a recent paper [1]. In particular, we started with an *universal approximation theorem* for DeepOnets. This theorem states that there exists a DeepOnet that can approximate any measurable (with respect to an underlying measure on the Banach space), operator mapping one Banach space into another. Thus, DeepOnets have analogous universal approximation property to deep neural networks for finite-dimensional functions.

However, this result does not provide any quantitative information on the size of the DeepOnet. In particular, it does not rule out the possibility that the DeepOnet might suffer from the *curse of dimensionality* i.e., exponential growth of the size of the DeepOnet can grow exponentially (or faster) with increasing accuracy. The main point of the talk and the paper [1] is to prove that DeepOnets can break this curse of dimensionality for operators that stem from PDEs.

To this end, we presented a detailed and careful analysis of different components of the DeepOnet approximation error and showed that the total network size for approximating PDE based Operators only grew polynomially with decreasing accuracy. This analysis was carried out for four representative examples, that of a forced pendulum, an elliptic PDE with variable coefficients and for nonlinear Parabolic Allen-Cahn equations and nonlinear hyperbolic scalar conservation laws.

References

[1] S. Lanthaler, S. Mishra and G. E. Karniadakis, *Error estimates for DeepOnets: A deep learning framework in infinite dimensions*, arXiv:2102.09618, 2021.

A Machine Learning Framework for High-Dimensional Mean Field Games and Optimal Control

Lars Ruthotto

(joint work with Derek Onken, Samy Wu Fung, Levon Nurbekyan, Xingjian Li, Stanley Osher)

We consider the numerical solution of Hamilton Jacobi Bellman (HJB) equations arising in mean field games and optimal control problems whose state space dimension is in the tens or hundreds. In this setting, most existing numerical methods that rely on spatial discretization (e.g., using grids or meshes) are affected by the curse of dimensionality (CoD); i.e., their computational complexity grows exponentially with the state space dimension.

To mitigate the CoD, our framework parameterizes the value function with a neural network that is designed specifically to allow accurate and efficient computation of first- and second-order derivative information. We train the neural network weights by minimizing the objective function of the original problem with additional penalties that enforce the HJB equations subject to neural ODE constraints. In mean field games, those constraints arise from the method of characteristics applied to the continuity equation that describes the evolution of the population density. Similarly, in general optimal control problems, the neural ODE constraint arises from Pontryagin's maximum principle and a closed-loop feedback form. A key benefit of our framework is that no training data is needed, e.g., no numerical solutions to the problem need to be computed before training and - once trained in an offline setting - the neural network can be evaluated quickly to produce approximately optimal policies.

We illustrate our approach and its efficacy using several numerical experiments. To show the framework's generality, we consider applications such as optimal transport, deep generative modeling, mean field games for crowd motion, and multi-agent optimal control.

References

[1] D. Onken, L. Nurbekyan, X. Li, S. W. Fung, S. Osher, and L. Ruthotto. *A Neural Network Approach for High-Dimensional Optimal Control*. arXiv, 2021.

[2] D. Onken, S. Wu Fung, X. Li, and L. Ruthotto. *OT-flow: Fast and accurate continuous normalizing flows via optimal transport.* In 35th Conference on AAAI, 2021.

[3] L. Ruthotto and E. Haber. *An Introduction to Deep Generative Modeling*. GAMM Mitteilungen, 2021.

[4] L. Ruthotto, S. J. Osher, W. Li, L. Nurbekyan, and S. W. Fung. *A machine learning framework for solving high-dimensional mean field game and mean field control problems.* Proceedings of the National Academy of Sciences, 117(17):9183 – 9193, 2020.

Very weak Space-Time Formulations and Fast Solvers

Karsten Urban

(joint work with Davide Palitta, Valeria Simoncini and Julian Henning)

During the last years there has been an increasing interest in space-time methods for time-dependent partial differential equations (PDEs). This has various aspects from the analysis concerning well-posedness of evolutionary problems, the construction and analysis of discretizations for determining numerical approximations up to the development and realization of efficient numerical solvers. Nowadays, there is a rich literature, a survey goes well beyond this abstract. However, also the notion *space-time* is used with different meanings and interpretations in the literature.

We start by a time-dependent linear PDE and derive a space-time *variational* formulation by multiplying with a test function in both space as well as time variables and integrate over these variables. Depending on the problem at hand, one performs integration by parts in order to bring certain derivatives onto the test functions.

What we are after, is a well-posed variational formulation of the following type: Given two Hilbert spaces, the trial space X and the test space Y, a bilinear form $b : X \times Y \to \mathbb{R}$ and a right-hand side $f \in Y' := \{\ell : Y \to \mathbb{R}, \ell \text{ linear}\}$, one seeks

$$u \in X \quad \text{such that} \quad b(u,v) = f(v) \quad \text{for all } v \in Y. \tag{1}$$

We are considering three classes of time-dependent PDEs here, namely

- the heat equation: $Lu := u_t - \Delta u = f$, $u(0) = u_0$,
- the linear transport problem: $Lu := u_t + \beta \cdot \nabla u = f$, $u(0) = u_0$,
- the wave equation: $Lu := u_{tt} - \Delta u = f$, $u(0) = u_0$, $u_t(0) = v_0$,

all on some domain $\Omega \subset \mathbb{R}^d$ in space and some time interval $I = (0,T)$. We want to develop formulations of type (1) for these three example classes such that

(a) the problem (1) is well-posed, i.e., existence, uniqueness and stability;

(b) we can construct a Petrov-Galerkin discretization consisting of finite-dimensional trial $X_\delta \subset X$ and test spaces $Y_\delta \subset Y$ such that $\dim X_\delta = \mathcal{N}_\delta < \infty$ allowing for uniform stability;

(c) the discrete problem

$$u_\delta \in X_\delta : \quad b(u_\delta, v_\delta) = f(v_\delta) \quad \text{for all } v_\delta \in Y_\delta$$

is well-posed, converges to u as $\mathcal{N}_\delta \to \infty$ and can be solved in an efficient manner, in particular as compared with standard time-marching schemes.

(a) Well-posed operator problem. With decent definitions of X and Y, it is not difficult to show in all mentioned problem classes that the arising bilinear form is bounded, i.e., there exists a constant $0 < C < \infty$ such that

$$|b(u,v)| \le C \, \|u\|_X \, \|v\|_Y, \quad u \in X, v \in Y.$$

Then, the famous Nečas theorem ensures that (1) admits a unique solution $u \in X$ such that $\|u\|_X \le c \, \|f\|_{Y'}$ *if and only if*

$$\text{(i)} \qquad \exists \beta > 0 : \qquad \sup_{v \in Y} \frac{b(u,v)}{\|v\|_Y} \ge \beta \, \|u\|_X \text{ for all } u \in X \qquad \text{(inf-sup stability);}$$

$$\text{(ii)} \qquad \text{for any } 0 \ne v \in Y \quad \exists u \in X : \quad b(u,v) \ne 0 \qquad \text{(injectivity).}$$

Moreover $c = \beta^{-1}$.

For the heat equation, well-posedness with $\beta = C = 1$ (i.e., optimal stability) was shown in [5, 6] for a “standard” space-time variational form. The situation changes for the linear transport equation. In fact, [2] presents an optimally stable *very-weak* variational form, i.e., using the bilinear form $b(u,v) := (u, L^* v)_{L_2(I \times \Omega)} = (u, -v_t - \beta \cdot \nabla v)_{L_2(I\times\Omega)}$, the trial space $X := L_2(I; L_2(\Omega)) = L_2(I \times \Omega)$ and some non-standard test space Y.

This idea has been extended for the wave equation in [4]. In fact, we set $X := L_2(I \times \Omega)$, $b(u,v) := (u, -v_{tt} + \Delta v)_{L_2(I \times \Omega)}$ and $Y := \operatorname{clos}_{\|\cdot\|_Y}\{v \in C^2(I \times \Omega) : v(T) = v_t(T) = 0, v_{|\partial\Omega} = 0\}$. We show that $\|v\|_Y := \|L^* v\|_{L_2(I \times \Omega)}$ is a norm on Y due to the following statement.

Theorem 1. *Let $u_0 \in L_2(\Omega)$, $u_1 \in H^{-1}(\Omega)$ and $f \in C([0,T]; H^{-1}(\Omega))$. Then, the problem $\ddot{w}(t) + A\,w(t) = f(t)$, $t \in (0,T)$, $w(0) = u_0, \dot{w}(0) = u_1$ admits a unique solution $w \in C^2([0,T]; H^{-2}(\Omega)) \cap C^1([0,T]; H^{-1}(\Omega)) \cap C([0,T], L_2(\Omega))$.*

This theorem also ensures well-posedness of the above very-weak variational formulation for the wave equation, [4].

(b) Petrov-Galerkin discretization. In order to obtain an unconditionally stable Petrov-Galerkin discretization, we need to verify the *LBB condition* i.e., the existence of a constant $\beta > 0$ independent of $\mathcal{N}_\delta \to \infty$ such that

$$\inf_{u_\delta \in X_\delta} \sup_{v_\delta \in Y_\delta} \frac{b(u_\delta, v_\delta)}{\|u_\delta\|_X \, \|v_\delta\|_Y} \geq \beta > 0. \tag{1.2}$$

In order to do so, we follow [1] and start by choosing the test space, here

$$\begin{aligned} R_{\Delta t} &:= \operatorname{span}\{\varrho^1, ..., \varrho^{N_t}\} \subset \{\varrho \in H^2(I) :\, \varrho(T) = \dot{\varrho}(T) = 0\} \\ Z_h &:= \operatorname{span}\{\phi_1, ..., \phi_{N_h}\} \subset H_0^1(\Omega) \cap H^2(\Omega) \end{aligned}$$

and setting $Y_\delta := R_{\Delta t} \otimes Z_h$. Then, defining the non-standard trial space as $X_\delta := L^*(Y_\delta)$ ensures optimal stability, i.e., $\beta = 1$. By *non-standard* we mean that X_δ is not a standard spline space, but arises as the image of a spline space Y_δ under the adjoint operator L^*. The advantage of this approach is that inf-sup stability does not need to be ensured by constructing specific test functions, but is automatically guaranteed.

(c) Efficient numerical solvers. In all three problem classes, using a space-time variational form yields a linear system of equations $\mathbf{B}_\delta \mathbf{u}_\delta = \mathbf{f}_\delta$, where the stiffness matrix is a sum of tensor product matrices. For the heat equation, we have reported in [3] that a matrix-oriented Sylvester-type solver can outperform standard time-stepping schemes (Crank-Nicolson) in terms of CPU-time.

For the wave equation, the situation is more involved. In fact, the stiffness matrix takes the form

$$\mathbf{B}_\delta = A_{\Delta t} \otimes M_h + N_{\Delta t} \otimes N_h^T + N_{\Delta t}^T \otimes N_h + M_{\Delta t} \otimes A_h,$$

where some of the involved matrices are singular. However, constructing a matrix-oriented low-rank Galerkin projection combined with a rational Krylov subspace method gives rise to a quite efficient numerical solver. We have observed in several numerical experiments that this space-time numerical method outperforms the Crank-Nicolson method[1] if the solution has only the minimal regularity, i.e., $u \in L_2(I \times \Omega)$, and not more. On the other hand, if the solution admits more regularity, then the quadratic convergence of the Crank-Nicolson method cannot be reached by such a low-order[2] space-time variational approach. Details can be found in [4].

[1] Crank-Nicolson for the wave equation is Newmark's method with $\beta = 1/4$ and $\gamma = 1/2$.

[2] Low-order due to the very-weak approach using piecewise constant trial functions.

References

[1] J. Brunken, K. Smetana, and K. Urban. *(Parametrized) First Order Transport Equations: Realization of Optimally Stable Petrov-Galerkin Methods.* SIAM J. Sci. Comput., 41(1):A592–A621, 2019.

[2] W. Dahmen, C. Huang, C. Schwab, and G. Welper. *Adaptive Petrov-Galerkin methods for first order transport equations.* SIAM J. Numer. Anal., 50(5):2420–2445, 2012.

[3] J. Henning, D. Palitta, V. Simoncini, and K. Urban. *Matrix Oriented Reduction of Space-Time Petrov-Galerkin Variational Problems.* In F. J. Vermolen and C. Vuik, eds., Numerical Mathematics and Advanced Applications ENUMATH 2019, pp. 1049–1057. Springer, 2019.

[4] J. Henning, D. Palitta, V. Simoncini, and K. Urban. *Very Weak Space-Time Variational Formulation for the Wave Equation: Analysis and Efficient Numerical Solution.* arXiv 2107.12119, 2021.

[5] K. Urban and A. T. Patera. *A new error bound for reduced basis approximation of parabolic partial differential equations.* C. R. Math. Acad. Sci. Paris, 350(3-4):203–207, 2012.

[6] K. Urban and A. T. Patera. *An improved error bound for reduced basis approximation of linear parabolic problems.* Math. Comp., 83(288):1599–1615, 2014.

Deep Learning in Numerical Analysis

Philipp Grohs

Several research questions that are not answered within the classical framework of learning theory are approached in the new field of mathematical analysis of deep learning.

References

[1] J. Berner, P. Grohs, G. Kutyniok and P. Petersen. *The Modern Mathematics of Deep Learning.* arXiv:2105.04026, 2021.

A Spectral Galerkin Method for the Solution of Reaction-Diffusion Equations on Metric Graphs

Anna Weller
(joint work with Mark Ainsworth)

The accumulation of intraneuronal *tau-tangles* is a hallmark of Alzheimer's Disease (AD). Tau proteins aggregate in the neurons in form of tangles, affecting neuronal function and leading to neuronal death. Whereas the tau-tangles in the brain of AD patients has been known for more than a century, it is a relatively recent hypothesis that they may travel from one neuron to another, inducing tangles in neighboring neurons in a prion-like fashion. Together with extraneuronal aggregation of beta-amyloid peptides, tau-tangles are believed to be an important factor in AD and other neurodegenerative disorders.

A major hurdle in interpretation of *in vivo* data is the analysis of the complex interplay between these multiple factors of pathology and the complexity of the brain network. To this end, it is highly temping to develop a *Global Brainsphere Model* for the simulation of AD [3]. Several data have been analyzed in preparation for this model to investigate, in particular, the evolution and effects of tau-tangles

[6]. The presumed prion-like spreading mechanism motivates the simulation of tau as a reaction-diffusion process on the brain network [5]. The numerical solution of such systems on a continuous interpretation of the network, a metric graph, is the objective of this work in progress [1].

A combinatorial graph $G = (V, E)$ consists of a vertex set V and a set of edges E, where an edge e_{ij} represents a connection between two vertices v_i and v_j. This discrete understanding of a graph can be extended to a continuous METRIC GRAPH Γ by identifying each edge e_{ij} with an interval $[0, \ell_{ij}]$, where ℓ_{ij} is the length of the edge. By this, the graph can be interpreted as topological space. Equipped with a differential operator $\mathcal{H}$, for example the negative second derivative, these metric graphs are often referred to as QUANTUM GRAPHS [2].

By the spectrum of a quantum graph we understand the spectrum of $\mathcal{H}$ on Γ, determined by the eigenvalue problem

$$\frac{\partial^2}{\partial x^2}\, u(x) = \lambda\, u(x) \tag{1}$$

on Γ with Neumann-Kirchhoff boundary conditions

$$\begin{cases} u(x) \text{ is continuous on } \Gamma \\ \displaystyle\sum_{e \in E_{v_i}} \frac{\partial u|_{ij}}{\partial x}\Big|_{x=0} = 0 \text{ for all } v_i \in V. \end{cases} \tag{2}$$

For graphs with uniform edge length ℓ, i.e. EQUILATERAL GRAPHS, there exists a well known relation between part of the spectrum of Γ and the underlying discrete graph G, first derived by [4]. Namely, for $\lambda \neq \left(\frac{k\pi}{\ell}\right)^2$, we have that $\lambda \in \sigma(\Gamma)$ if and only if $(1 - \cos(\sqrt{\lambda})\ell) \in \sigma(\Delta_G)$, where Δ_G is the HARMONIC LAPLACIAN acting as $(\Delta_G u)(v_i) = u(v_i) - \frac{1}{\deg(v_i)} \sum_{v_j \sim v_i} u(v_j)$. We deduce that for an eigenpair (μ, ϕ) of Δ_G, the following functions defined on each edge

$$\varphi|_{ij}(x; \lambda_k) = \frac{1}{\sin(\sqrt{\lambda_k}\ell)} \Big(\phi(v_i)\sin(\sqrt{\lambda_k}(\ell - x)) + \phi(v_j)\sin(\sqrt{\lambda_k}x)\Big) \tag{3}$$

where

$$\lambda_k = \begin{cases} \frac{1}{\ell}(\arccos(1-\mu) + k\pi)^2 & \text{for } k \text{ even} \\ \frac{1}{\ell}(\arccos(1-\mu) - (k+1)\pi)^2 & \text{for } k \text{ odd.} \end{cases} \tag{4}$$

are eigenfunctions of Γ. We call these type of eigenfunctions VERTEX EIGENFUNCTIONS, as they can be completely determined by the values of the eigenvectors of a discrete matrix defined on the vertices of G.

For the remaining NON-VERTEX EIGENVALUES, $\varphi|_{ij}$ in (3) is not well defined. To construct the corresponding eigenfunctions, we make use of a simple observation by [2], that the elimination of a vertex of degree two by combining the two adjacent edges into one edge does not change the solutions of the eigenvalue problem under the given boundary conditions. This also means that we can add k artificial vertices of degree two at each edge to create an *extended graph* $\tilde{\Gamma}_k$ with edge length $\tilde{\ell}_k = \frac{\ell}{k+1}$. By this, we can construct the eigenfunctions of $\lambda = \left(\frac{k\pi}{\ell}\right)^2$ by applying formula (3) and (4) to the extended graph as $\varphi|_{ij}$ now is well defined on the edges of the extended graph. Together, we obtain an increasing sequence of vertex and non-vertex eigenvalues with corresponding eigenfunctions.

We now consider the reaction-diffusion equation

$$\frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} = f(u(x,t)) \tag{5}$$

on a metric graph with Neumann-Kirchhoff boundary conditions (2) and initial condition $u(x,0) = u_0$. The weak formulation is given by: Find $u \in H^1(\Gamma)$ with

$$\begin{aligned} \frac{\partial}{\partial t}(u,\phi) + \left(\frac{\partial u}{\partial x}, \frac{\partial \phi}{\partial x}\right) &= (f(u(x,t)), \phi)\,, \\ (u(x,0),\phi) &= (u_0(x),\phi) \quad \forall \phi \in H^1(\Gamma), \end{aligned}$$

where the inner product is defined by

$$(u,w)_\Gamma = \sum_{e \in E} \int_e u(x)w(x)dx.$$

Let now $\Lambda_k := \left(\frac{k\pi}{\ell}\right)^2$ and $X_k := \operatorname{span}\{\varphi_\lambda : \lambda \leq \Lambda_k\}$, where φ_λ is the eigenfunction to eigenvalue λ. Then, the spectral Galerkin approximation consists of finding $u_k \in X_k$ with

$$\frac{\partial}{\partial t}(u_k,\varphi_\lambda) + \left(\frac{\partial u_k}{\partial x}, \frac{\partial \varphi_\lambda}{\partial x}\right) = (f(u_k(x,t)), \varphi_\lambda) \quad \forall \varphi_\lambda \in X_k. \tag{6}$$

Representing u_k as $u_k = \sum_{\lambda \leq \Lambda_k} c_\lambda \varphi_\lambda$ for constants c_λ reduces this to the ordinary differential equation

$$\frac{\partial}{\partial t}(e^{\mathbf{\Lambda} t}\mathbf{c}) = e^{\mathbf{\Lambda} t}(f(u(x,t)), \Phi),$$

where $\mathbf{\Lambda}$ is a diagonal matrix containing the eigenvalues λ and $\mathbf{c}$ a vector containing the constants c_λ. By this, we derived a semidiscretization of a reaction-diffusion equations on a metric graph which now can be solved by a suitable method for ordinary differential equations.

References

[1] M. Ainsworth, A. Weller. *A Spectral Galerkin Method for the Solution of Partial Differential Equations on Metric Graphs*, *in preparation.*

[2] G. Berkolaiko and P. Kuchment. *Introduction to Quantum Graphs*, American Mathematical Society, Providence, RI, 2013.

[3] A. Kunoth, Y. Shao, A.Weller. *A Computational Brainsphere Model for the Simulation of Alzheimer's Disease*, *in preparation.*

[4] K. Pankrashkin. *Spectra of Schrödinger Operators on equilateral Quantum Graphs*, Lett. Math. Phys., 77(2):139–154, 2006.

[5] A. Raj, A. Kuceyeski, M. Weiner. *A Network Diffusion Model of Disease Progression in Dementia*, Neuron, 73(6):1204–15, 2012.

[6] A. Weller, G. N. Bischof, P. Schlueter, N. Richter, J. Dronse, O. Onur, B. Neumaier, J. Kukolja, K. Langen, G. Fink, A. Kunoth, Y. Shao, T. van Eimeren, A. Drzezga, *Finding New Communities: A Principle of Neuronal Network Reorganization in Alzheimer's Disease*, Brain Connectivity, 11(3):225–238, 2021.

Neural Network Approximation of Functions related to PDEs

Philipp Petersen

(joint work with Andrei Caragea, Fabian Laakmann, Carlo Marcati, Joost Opschoor, Christoph Schwab, Felix Voigtlaender)

In this talk, we discussed approximation theoretical aspects of the mathematical theory of deep learning [2]. Concretely, we focus on functions that exhibit structured singularities.

The first type of singularities that we studied are the following: Let $d \in \mathbb{N}$ and let $\mathcal{C}_1 \subset L^\infty([0,1]^d)$, $\mathcal{C}_2 \subset L^\infty([0,1]^{d-1})$ be two function classes. We study functions of the form $f = f_1 + \chi_B f_2$, where $f_1, f_2 \in \mathcal{C}_1$ and $B \subset [0,1]^d$ is a set such that ∂B can locally be parametrised by a function in $\mathcal{C}_2$. We call functions of this form *functions with structured singularities of type* $\mathcal{C}_1, \mathcal{C}_2$. These functions are potentially high dimensional functions that are generally not continuous. However, we expect that these functions can still be very efficiently approximated if the singularity curve can be properly resolved, which we demonstrate for a number of examples.

The first result that we presented is that if $\mathcal{C}_1 = C^k([0,1]^d)$, $\mathcal{C}_2 = C^k([0,1]^{d-1})$, then for every function f with structured singularities of type $\mathcal{C}_1, \mathcal{C}_2$ and every $\epsilon > 0$ there exists a neural network Φ satisfying

$$\|f - \Phi\|_{L^2([0,1]^d)} \leq \epsilon$$

and Φ has not more than $\epsilon^{-2(d-1)/k}$ non-zero entries [7]. *This implies that neural networks can approximate discontinuous functions very efficiently as long as the singularity curve is sufficiently smooth.*

One issue with the result above is that it suffers from the so-called *curse of dimensionality*. In this context, we say that an approximation method suffers from the curse of dimensionality if the approximation rate deteriorates exponentially with increasing dimensions. However, in practical applications of deep neural networks the input dimensions are immense and one typically does not observe a deterioration with higher dimensions.

A famous class of functions for which the curse of dimensionality can be overcome by approximation through neural networks is the *Barron class* [1]. These are functions $f\colon \mathbb{R}^d \to \mathbb{R}$ such that for a constant $c > 0$

$$f(x) = c + \int_{\mathbb{R}^d} (e^{i\langle x, \xi\rangle} - 1) F(\xi) d\xi, \text{ for } x \in \mathbb{R}^d, \text{ where } \int_{\mathbb{R}^d} |\xi||F(\xi)| d\xi < \infty.$$

Functions in the Barron type can be approximated by neural networks with N neurons to an L^2 error of $N^{-1/2}$ on the unit ball. Interestingly, the approximation rate, while slow, is independent of the underlying dimension. In [3], we showed that this result implies that for functions f with structured singularities of type $\mathcal{C}_1, \mathcal{C}_2$, where $\mathcal{C}_1$ contains only constant functions and $\mathcal{C}_2$ is the set of functions of Barron type it holds that there exists a neural network Φ such that for every probability measure μ

$$\mu(\{x \in \mathbb{R}^d \colon f(x) \neq \Phi(x)\}) \lesssim d^{3/2} N^{-\alpha/2},$$

where α depends on μ, [3]. For measures with bounded densities, we have that $\alpha = 1$. *As a result, we observe that neural networks have the surprising property of approximating without the curse of dimension arbitrarily high dimensional functions with non-trivial discontinuities.*

Another instance of neural networks resolving potentially complicated discontinuities to achieve very high approximation rates of non-smooth functions was found in the context of transport equations, [4]. We consider *parametric transport equations*

$$\begin{aligned}\partial_t u(t,x,\eta) + V(t,x,\eta) \cdot \nabla_x u(t,x,\eta) &= 0,\\ u(0,x,\eta) &= u_0(x),\end{aligned}$$

where $t \in [0,T]$ is a *time parameter*, $x \in \mathbb{R}^n$ is a *spatial coordinate*, and $\eta \in [0,1]^D$ is a *parameter* for some $n, D \in \mathbb{N}$, and $T > 0$. The vector field $V \in C^k([0,T] \times \mathbb{R}^n \times [0,1]^D; \mathbb{R}^n)$ and the initial condition $u_0 \in C^s(\mathbb{R}^n; \mathbb{R})$ are given with $s, k \in \mathbb{N}$. In this case, we can show that the characteristic curves of u are as smooth as the vector field V. Hence, if u_0 is a piecewise smooth function, then it follows by the method of characteristics that u has structured singularities along smooth curves.

Finally, we studied a slightly different type of structured singularities that arise as boundary effects in boundary value problems with analytic coefficients. In these and other related partial differential equations, it can be shown that the solutions are so-called weighted analytic functions. These functions are analytic in the interior of the domain but exhibit a controlled blow-up in their derivatives close to the boundary and the corners of the domains. By demonstrating that deep neural networks can re-approximate *hp*-finite element methods on general polyhedral domains, we demonstrate that deep neural networks can again successfully resolve complex singularities, [6, 5]. In this case, this leads to *exponential approximation rates of non-smooth functions* of neural networks with respect to the number of parameters of the networks.

References

[1] A. R. Barron, *Universal Approximation Bounds for Superpositions of a Sigmoidal Function,* IEEE Transactions on Information Theory, 39(3), 1993.

[2] J. Berner, P. Grohs, G. Kutyniok and P Petersen, *The modern mathematics of deep learning,* arXiv:2105.04026, 2021.

[3] A. Caragea, P. Petersen and F. Voigtlaender, *Neural network approximation and estimation of classifiers with classification boundary in a Barron class*, arXiv:2011.09363, 2020.

[4] F. Laakmann and P. Petersen. *Efficient approximation of solutions of parametric linear transport equations by ReLU DNNs,* Advances in Computational Mathematics, 47(1):1–32, 2021.

[5] C. Marcati and J. A. A. Opschoor and P. Petersen and C. Schwab, *Exponential ReLU Neural Network Approximation Rates for Point and Edge Singularities*, arXiv:2010.12217, 2020.

[6] J. A. A. Opschoor, P. Petersen and C. Schwab. *Deep ReLU networks and high-order finite element methods,* Analysis and Applications, 18:05, 715–770, 2020.

[7] P. Petersen and F. Voigtlaender. *Optimal approximation of piecewise smooth functions using deep ReLU neural networks,* Neural Networks, 108:296–330, 2018.

Numerical Solution of Hamilton Jacobi Bellman Equation and further non-linear high dimensional PDE's

Reinhold Schneider

The Hamilton-Jacobi-Bellman (HJB) equation associated to infinite horizon optimal control problems is non linear and suffers from the curse of dimensionality. Low rank hierarchical tensor product approximations are used to solve the operator equations resulting from the reduction of the HJB to a sequence of linear, hyperbolic PDEs.

References

[1] M. Oster, L. Sallandt, R. Schneider, *Approximating the Stationary Bellman Equation by Hierarchical Tensor Products*, arXiv:1911.00279, 2019.

Low-Rank Tensor-Based Transport for high-dimensional Bayesian Inference

Robert Scheichl
(joint work with Karim Anaya-Izquierdo, Tiangang Cui, Sergey Dolgov, Colin Fox, Lars Grasedyck and Paul Rohrbach)

General multivariate distributions are notoriously expensive to sample from, particularly the high-dimensional posterior distributions in PDE-constrained inverse problems. In this talk, I present a joint paper with K. Anaya-Izquierdo, S. Dolgov and C. Fox [1] on a sampler for arbitrary continuous multivariate distributions. The approach is based on low-rank surrogates in the tensor-train (TT) format, a methodology that has been exploited for many years for scalable, high-dimensional density function approximation in quantum physics and chemistry.

We build upon recent developments of the cross approximation algorithms in linear algebra to construct a tensor-train approximation to the target probability density function using a small number of function evaluations [7, 6]. For sufficiently smooth distributions the storage required for accurate tensor-train approximations is moderate, scaling linearly with dimension. In turn, the structure of the tensor-train surrogate allows sampling by an efficient conditional distribution method since marginal distributions are computable with linear complexity in dimension. Expected values of non-smooth quantities of interest, with respect to the surrogate distribution, can be estimated using transformed independent uniformly-random seeds that provide Monte Carlo quadrature, or transformed points from a quasi-Monte Carlo lattice to give more efficient quasi-Monte Carlo quadrature. Unbiased estimates may be calculated by correcting the transformed random seeds using a Metropolis–Hastings accept/reject step, while the quasi-Monte Carlo quadrature may be corrected either by a control-variate strategy, or by importance weighting.

We show that the error in the tensor-train approximation propagates linearly into the Metropolis–Hastings rejection rate and the integrated autocorrelation time of the resulting Markov chain; thus the integrated autocorrelation time may be made arbitrarily close to , implying that, asymptotic in sample size, the cost per effectively independent sample is one target density evaluation plus the cheap tensor-train surrogate proposal that has linear cost with dimension.

As an exemplary problem, the methods are demonstrated on a PDE-constrained inverse diffusion problem. The delayed rejection adaptive Metropolis (DRAM) algorithm [3] is used as a benchmark. In all computed examples, the importance-weight corrected quasi-Monte Carlo quadrature performs best, and is more efficient than DRAM by orders of magnitude across a wide range of approximation accuracies and sample sizes. Indeed, all the methods developed here significantly outperform DRAM in all computed examples.

In this talk, I will also highlight the link to transport-based sampling algorithms for high-dimensional distributions, in the spirit of [5], and to normalizing flows [8] in machine learning. In particular, the method falls square into the category of Knothe-Rosenblatt rearrangment-based triangular transport maps proposed by Marzouk, Moshely, Parno and Spantini [4], when the low-rank tensor-train format is used as the model class for approximating the transport.

As a starting point for a full theoretical analysis, in a recently submitted preprint with Paul Rohrbach, Sergey Dolgov and Lars Grasedyck [9], we were able to also give rigorous a priori bounds on the necessary ranks to approximate general multivariate Gaussian distributions in the functional tensor-train representation. It is shown that under suitable conditions on the precision matrix, the Gaussian density can be approximated to high accuracy without suffering from an exponential growth of complexity as the dimension increases. In fact, the growth with the dimension is polylogarithmic. These results provide a rigorous justification of the suitability and the limitations of low-rank tensor methods in a simple but important model case. Numerical experiments confirm that the rank bounds capture the qualitative behavior of the rank structure when varying the parameters of the precision matrix and the accuracy of the approximation.

Finally, I will highlight extensions of the approach to tackle more strongly concentrating probability distributions, as common in Bayesian inverse problems, via a multi-layered approach. In [2], Tiangang Cui and Sergey Dolgov recently were able to extend the TT-sampling algorithm to a Deep Inverse Rosenblatt Transport (DIRT) algorithm that achieves the approximation of the transport in a convolutional setting with significantly lower rank approximations of the transport maps between the individual bridging densities.

References

[1] K. Anaya-Izquierdo, S. Dolgov, C. Fox and R. Scheichl, *Approximation and sampling of multivariate probability distributions in the tensor train decomposition*, Stat. Comput. **30** (2020), 603–625.

[2] T. Cui and S. Dolgov, *Deep composition of tensor trains using squared inverse Rosenblatt transports*, Preprint arXiv:2007.06968 (2020),

[3] H. Haario, M. Laine, A. Mira, and E. Saksman, *DRAM: Efficient adaptive MCMC*, Stat. Comput. **16** (2006), 339–354.

[4] Y. Marzouk, T. Moselhy, M. Parno and A. Spantini *Sampling via measure transport: An introduction* in Handbook of Uncertainty Quantification (R.. Ghanem, D. Higdon, H. Owhadi, eds.), Springer (2017), pp. 1-41.

[5] T.A. El Moselhy and Y.M. Marzouk, *Bayesian inference with optimal maps*, J. Comput. Phys. **231** (2012), 7815–7850.

[6] I. Oseledets, *Constructive representation of functions in low-rank tensor formats*, Constr. Approx. **37** (2013), 1–18.

[7] I. Oseledets and E. Tyrtyshnikov, *TT-cross approximation for multidimensional arrays*, Linear Algebra Appl. **432** (2010), 70–88.

[8] D. Rezende and S. Mohamed, *Variational inference with normalizing flows*, in Proceedings of the 32nd International Conference on Machine Learning, PMLR **37** (2015), 1530–1538.

[9] P.B. Rohrbach, S. Dolgov, L. Grasedyck and R. Scheichl, *Rank bounds for approximating Gaussian densities in the tensor-train format*, Preprint arXiv:2001.08187 (2020), 1–25.

Approximation power of neural networks

Josiah Park

(joint work with Ingrid Daubechies, Ronald DeVore, Nadav Dym, Shira Faigenbaum-Golovin, Shahar Z. Kovalsky, Kung-Ching Lin, Guergana Petrova, Barak Sober)

In the desire to quantify the success of neural networks in deep learning and other applications, there is a great interest in understanding which functions are efficiently approximated by the outputs of neural networks. By now, there exists a variety of results which show that a wide range of functions can be approximated with sometimes surprising accuracy by these outputs. For example, it is known that the set of functions that can be approximated with exponential accuracy (in terms of the number of parameters used) includes, on one hand, very smooth functions such as polynomials and analytic functions (see e.g. [2, 4, 5]) and, on the other hand, very rough functions such as the Weierstrass function (see e.g. [1, 3]), which is nowhere differentiable. In this talk, we add to the latter class of rough functions by showing that it also includes refinable functions. Namely, we show that refinable functions are approximated by the outputs of deep ReLU networks with a fixed width and increasing depth with accuracy exponential in terms of their number of parameters. Our results apply to functions used in the standard construction of wavelets as well as to functions constructed via subdivision algorithms in Computer Aided Geometric Design.

References

[1] I. Daubechies, R. DeVore, S. Foucart, B. Hanin, G. Petrova, *Nonlinear approximation and (deep) relu networks*, arXiv, Constructive Approximation, to appear.

[2] W. E, Q. Wang, *Exponential convergence of the deep neural network approximation for analytic functions*, Sci. China Math **61** (2018), 1733–1740.

[3] D. Elbrachter, D. Perekrestenko, P. Grohs, H. Bölcskei, *Deep Neural Network Approximation Theory*, IEEE Transactions of Information Theory, **67**(5) (2021), 2581–2623.

[4] J.A.A. Opschoor, Ch. Schwab, J. Zech , *Exponential ReLU DNN Expression of Holomorphic Maps in High Dimension*, Constructive Approximation (2021), doi.org/10.1007/s00365-021-09542-5.

[5] D. Yarotsky, *Error bounds for approximations with deep relu networks*, Neural Networks **94** (2017), 103–114.

Modeling and Learning of X-ray Microscopy Data

Peter Binev

(joint work with Kelsey Larkin, Zineb Saghi, Toby Sanders)

Processing of hyperspectral data, and Energy Dispersive X-ray spectroscopy (EDX) in particular, is challenging especially in the data-poor situation in which the data is not enough to recover the complete spectrum. We present a general approach to processing hyperspectral data that was tested on EDX tomography data. In this experiment a specimen was observed from 37 different tilt angles α by scanning it at a rectangular array of positions (x, y) using a focussed electron beam. At each position $p = p(\alpha, x, y)$ the beam is disturbing the atoms on its way and some of them loose a lower-orbital electron that is then replaced by a higher-orbital electron emiting an X-ray which energy is specific for the atom and corresponds to the difference of the energies required to be at each of the two orbitals. The spectrum is usually discretized into thousands of energy levels, e.g. $D = 4000$, and the EDX data at p is a D-dimensional vector $s(p) = \big(s_t(p)\big)_{t=1}^{D} \in \mathbb{Z}_+^D$ of counts of the detected X-rays at each energy level. Typically, only a total of a few hundreds of X-rays are detected per position p. Therefore, the data is insufficient to represent phenomena in the entire $\mathbb{Z}_+^D$ and a reliable processing of it is possible only if there is a sparse representation of $s(p)$ that contains the information of interest.

Modeling of EDX data. We consider $s(p)$ as a discretized realization of a random variable with a probability density function $f_p(t) = b(t) + \sum_{k=1}^{n} a_k \varphi_k(t)$ that is a linear combination of the probability density functions φ_k representing different emission lines of the atoms contained in the region of the specimen corresponding to the position $p = p(\alpha, x, y)$. The background signal $b(t)$ represents the phenomenon of background radiation that is present in the entire spectrum and is treated as a noise component with approximately known behavior. The constants a_k depend on the relative concentration of atoms in the region from the corresponding chemical element and the probability of emitting at the particular energy line. The random variable related to φ_k has a distribution close to Gaussian with mean μ_k and standard deviation σ_k. It can be approximated as $\varphi_k(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{t-\mu_k}{\sigma_k})^2}$ but more precise calculations require better models. For example, σ_k could be replaced with $\sigma_k + \gamma_k(t - \mu_k)$ especially for low mean energies μ_k or $\varphi_k(t)$ could be a linear combination of two or more Gaussian functions in case of very close energy lines of the same atom that are better to be modeled together. The values of μ_k and σ_k depend to the nature of the detector and its calibration and have to be estimated for each particular experiment. Thus, the functions φ_k can be considered as known in general but with a few parameters that have to be adjusted based on the data.

The problem of processing the EDX data is then reduced to solving the inverse problem of estimating the coefficients $a_k = a_k(p)$ given $s(p)$. It can be formulated as learning of the transform $T : \mathbb{Z}_+^D \to \mathbb{R}_+^n$ that for a given s finds $T(s) \approx \big(a_k\big)_{k=1}^n$. The amount of available data determines the precision of the estimates of the coefficients a_k with high probability. The standard approach uses only the top emission line per chemical element and determines the coefficients a_k as the relative amount of X-rays detected in the discrete energy levels that overlaps with the interval around μ_k, e.g. $[\mu_k - \sigma_k, \mu_k + \sigma_k]$, in which it is most likely that they were emitted by atoms representing this chemical element. While such an approach would provide the best estimate for an individual φ_k, it can be argued that a method using all the emission lines per chemical element should be able to provide a much better precision estimate. Another line of improvement is to base the estimates on regression rather than classification used in the standard approach.

Learning of the probability distribution. To approximate the probability distributions at each position p, we have to approximate the functions φ_k and estimate the counts $q_k(p)$ of the X-rays of $s(p)$ corresponding to each of them in order to estimate $a_k(p)$. The combined EDX spectrum at all positions $p = p(\alpha, x, y)$ gives the information needed to find all the emission lines contributing significant amounts of X-rays and to estimate the parameters of the corresponding functions φ_k, as well as the background signal b. In a small interval I, $b = \varphi_0$ can be approximated by a linear function in absence of significant absorption. Alternatively, it can be estimated as the difference of the combined spectrum and its approximation via a linear combination of φ_k, $k \in J, k > 0$ on I, in case some insignificant sources of X-rays are ignored. The functions $\varphi_k(t)$ are well localized and take significant values only in a small interval around μ_k. However, these intervals may overlap for different $\varphi_k(t)$. To simplify the analysis, we can identify the subintervals I of the (discrete) domain for t that represent groups of $\varphi_k(t)$ with overlapping essential intervals. Let us consider one such subinterval I and assume that the index set J represents the indices k in its group of $\varphi_k(t)$ together with the background function $b(t)$ which is denoted also by $\varphi_0(t)$ to simplify the exposition. One way of estimating the counts $q_k(p)$ for $k \in J$ is to define filters $\Lambda_k = \big(\lambda_{k,t}\big)_{t\in I}$ that are biorthogonal in expectation to the system of $\big(\varphi_k\big)_{k\in J}$. In addition to the biorthogonality condition

$$\langle \varphi_k, \Lambda_\ell \rangle_I := \sum_{t\in I} \varphi_k(t)\lambda_{\ell,t} = 0 \quad \text{for } k \neq \ell \,,$$

we require $\sum_{\ell\in J} \lambda_{\ell,t} = 1$ for every $t \in I$ to conserve the total X-ray count. To increase stability, we choose the filter Λ_ℓ that minimizes $\sum_{\ell\in J}\sum_{t\in I} |\lambda_{\ell,t}|^2$. The immediate estimate of $q_k(p)$ is $\tilde{q}_k(p) := \langle s(p), \Lambda_\ell \rangle_I$. However, due to the small amount of data these estimates could be very inaccurate and even some of $\tilde{q}_k(p)$ could be negative. Our discrete optimization procedure assigns initially a nonnegative integer value to $q_k(p)$ slightly underestimating $\tilde{q}_k(p)$ if it is positive. Then the values of

$q_k(p)$ for some p are increased by 1 until $\sum_{k\in J} q_k(p) = \sum_{k\in J} \tilde{q}_k(p) = \sum_{t\in I} s_t(p)$, as expected. Simultaneously, we also monitor the accumulated values of $q_k(p)$ for large sets of p, e.g. the frames F_α consisting of all $p = p(\alpha, x, y)$ for a fixed α and arbitrary x and y. Then, we require in addition that $\sum_{p\in F_\alpha} q_k(p) < 1+\sum_{p\in F_\alpha} \tilde{q}_k(p)$. The process of increasing the values of $q_k(p)$ is governed on a priority queue usually based on the current value of $\tilde{q}_k(p) - q_k(p)$ but may include other criteria. After the quantities $q_k(p)$ are estimated, we combine together the ones corresponding to the same chemical element to obtain the vector Q_z of relative concentrations $Q_z(p)$ of an element z at the position p.

EDX tomography. The goal is to determine the local concentrations of the chemical elements in the observed specimen. We use regularized minimization to perform the tomographic reconstruction. Let the volume of the specimen is divided into voxels v and let assume that $c_v(z)$ is the concentration of the element z at v and $c(z)$ is the vector of all the concentrations in the volume. We define a matrix $A = A(z)$ with entries $A_{p,v}$ that are the weights in the linear combination of $c(z)$ that describes the distribution of the element z in the result obtained at position p. Then the $Ac(z)$ should fit well Q_z. We use the volume based total variation of $c(z)$ as regularizer and solve

$$\min_{c(z)} \frac{\gamma}{2}\|Ac(z) - Q_z\|_2^2 + \|c(z)\|_{\mathrm{TV}}$$

using the alternating direction method of multipliers [1] with the openly available software [3]. The calculation of $A_{p,v}$ is usually based on the volume of interaction of the voxel v with the electron beam that represents p. This requires an alignment of the coordinate systems created for different tilt angles α since after each tilt the electron microscope has to be re calibrated. We adapt the center-of-mass approach of [4] for this alignment.

The proposed approach allows additional improvements of the estimates for Q_z that can be realized through the priority queue by iteratively modifying it based on the mismatches in the current calculation of $Ac(z) - Q_z$. It is also applicable to the situations of significant absorption of X-rays of higher energies by some of the atoms of the specimen. Then the matrix A becomes dependent on z and its entries are modified iteratively accounting for the calculated concentrations of the atoms on the way between the position of the beam and the EDX detectors.

A detailed description of the proposed methodology for processing hyperspectral data together with several examples and results is provided in [2].

Acknowledgements. This work was supported in part by the NSF grant DMS1720297, the University of South Carolina Magellan Scholar Program and Office of the Provost's Internal Grant Programs.

References

[1] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, *Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers* Foundations and Trends in Machine Learning, **3** (2011), 1–122

[2] K. M. Larkin, *Improved filtering of electron tomography EDX data*, Senior thesis, University of South Carolina (May 2020), URL `https://scholarcommons.sc.edu/senior_theses/351`

[3] T. Sanders, *MATLAB Imaging algorithms: Image Reconstruction, Restoration, and Alignment, with a Focus in Tomography*, (2017), DOI: `10.13140/RG.2.2.33492.60801`

[4] T. Sanders, M. Prange, C. Akatay, and P. Binev. *Physically motivated global alignment method for electron tomography*, Advanced Structural and Chemical Imaging, **1**:4 (2015)

Spline Insights to high dimensional Deep Learning Flows over Rough Boundaries

Richard Baraniuk

Spline functions and operators build a bridge between approximation theory and deep networks since a large class of deep networks can be written as a composition of max-affine spline operators.

References

[1] R. Balestriero, R. Baraniuk, *A spline theory of deep learning,* International Conference on Machine Learning (2018), 374-383.

[2] R. Balestriero, R. Baraniuk, *Mad Max: Affine Spline Insights into Deep Learning, arXiv:1805.06576 (2018).*

The Universal Approximation Theorem for complex-valued Neural Networks

Felix Voigtlaender

An important question in the theory of neural networks regards *universality*; that is, the ability of neural networks to approximate any given continuous function arbitrarily well (locally uniformly). Here, usually networks of fixed depth but arbitrarily large width are considered. Whether universality holds depends crucially on the chosen *activation function*; the question of which activation functions give rise to universal network classes has received significant interest [2, 3, 4, 8]. The most general result in this direction, presented in [8], states the following:

Theorem 1. *Let $\sigma : \mathbb{R} \to \mathbb{R}$ be locally bounded and assume that the closure of the set $\{x \in \mathbb{R} \colon \sigma \text{ discontinuous at } x\}$ is a null-set.*

Then for each $d \in \mathbb{N}$, the set $\mathcal{NN}_\sigma^d$ of shallow neural networks with d-dimensional input and activation function σ is universal if and only if *σ does not coincide (almost everywhere) with a polynomial. Here, universality means that for each continuous $f : \mathbb{R}^d \to \mathbb{R}$, there exists a sequence $(\Psi_n)_{n\in\mathbb{N}} \subset \mathcal{NN}_\sigma^d$ satisfying $\Psi_n \to f$ with locally uniform convergence.*

Remark. The claim remains true for the set $\mathcal{NN}_{\sigma,L}^d$ of *deep* neural networks with fixed number $L \in \mathbb{N}$ of hidden layers. For sufficiency, this follows since one can approximate the target function in the first layer and the identity functions in the subsequent layers. For necessity, one can show that if σ agrees almost everywhere with a polynomial p, then every $\Psi \in \mathcal{NN}^d_{\sigma,L}$ agrees (almost everywhere) with a polynomial of fixed degree, depending only on L and the degree of p.

In recent years, due to the impressive empirical success of deep neural networks in machine learning applications ("Deep Learning", see [7]), the universality properties of more special classes of neural networks have been analyzed in detail. For instance, [14] studies the universality of *convolutional* neural networks, while [9] considers so-called *residual networks.*

In this talk, I presented my recent work [12], in which I completely characterize those complex activation functions $\sigma : \mathbb{C} \to \mathbb{C}$ for which the associated network classes are universal. Such *complex-valued neural networks (CVNNs)* have received increased attention in recent years [10, 11, 13]. In particular, CVNNs have empirically been shown to provide increased stability of recurrent neural networks [13] and to outperform real-valued networks for problems in which the input is naturally complex-valued [11]. This superior performance is attributed in [11] to the ability of CVNNs to faithfully handle the phase of complex numbers.

Formally, a complex-valued neural network with activation function σ is a function

$$\Psi : \mathbb{C}^d \to \mathbb{C} \quad \text{of the form} \quad \Psi = T_L \circ (\sigma \circ T_{L-1}) \circ \cdots \circ (\sigma \circ T_0),$$

where $L \in \mathbb{N}$ denotes the number of hidden layers and each $T_\ell : \mathbb{R}^{N_{\ell+1}} \to \mathbb{R}^{N_\ell}$ is affine-linear (i.e., $T_\ell\, x = A_\ell\, x + b_\ell$ with $A_\ell \in \mathbb{R}^{N_{\ell+1} \times N_\ell}$ and $b_\ell \in \mathbb{R}^{N_{\ell+1}}$) and finally σ acts coordinatewise on vectors. Note that $N_0 = d$ and $N_{L+1} = 1$. Denoting by $\mathcal{NN}^d_{\sigma,L}$ the set of complex-valued networks with input dimension d, activation function σ and L hidden layers, the two main results from [12] state the following:

Theorem 2. *[Universal approximation; shallow case; see* [12, Theorem 1.3]*]*

Let $\sigma : \mathbb{C} \to \mathbb{C}$ be locally bounded and assume that the closure of the set $\{z \in \mathbb{C} \colon \sigma$ discontinuous at $z\}$ is a null-set. Let $d \in \mathbb{N}$ be arbitrary.

The set $\mathcal{NN}^d_{\sigma,1}$ is universal (in the sense that for any continuous $f : \mathbb{C}^d \to \mathbb{C}$ there exists a sequence $(\Psi_n)_{n\in\mathbb{N}} \subset \mathcal{NN}^d_{\sigma,1}$ satisfying $\Psi_n \to f$ locally uniformly) if and only if σ is not almost polyharmonic. *Here, we say that σ is almost polyharmonic if there exist a smooth function $\tau : \mathbb{C} \to \mathbb{C}$ and some $m \in \mathbb{N}$ satisfying $\sigma = \tau$ almost everywhere and $\Delta^m \tau \equiv 0$, with the Laplace operator Δ on $\mathbb{C} \cong \mathbb{R}^2$.*

Theorem 3. *[Universal approximation; deep case; see* [12, Theorem 1.4]*]*

Let $\sigma : \mathbb{C} \to \mathbb{C}$ be locally bounded and assume that the closure of the set $\{z \in \mathbb{C} \colon \sigma$ discontinuous at $z\}$ is a null-set. Furthermore, assume that none *of the following properties hold:*

a) we have $\sigma(z) = p(z, \overline{z})$ for almost all $z \in \mathbb{C}$, where p is a complex polynomial of two variables,

b) we have $\sigma = g$ almost everywhere or $\sigma = \overline{g}$ almost everywhere, where $g \colon \mathbb{C} \to \mathbb{C}$ is an entire holomorphic function.

Then, for each $L \in \mathbb{N}_{\geq 2}$ and each $d \in \mathbb{N}$, the class $\mathcal{NN}^d_{\sigma,L}$ of deep complex-valued neural networks with activation function σ and L hidden layers is universal.

Conversely, if $\sigma : \mathbb{C} \to \mathbb{C}$ is continuous and satisfies a) or b), then $\mathcal{NN}^d_{\sigma,L}$ is not *universal for any $d, L \in \mathbb{N}$.*

Remark.

- In dimension $d = 1$, it is relatively easy to see that continuous functions σ satisfying conditions *a)* or *b)* cannot be universal. Namely, if σ is holomorphic or anti-holomorphic, then every function in $\mathcal{NN}^1_{\sigma,L}$ will be holomorphic or anti-holomorphic as well, depending on whether L is even or odd. Since (anti)-holomorphicity is preserved under locally uniform limits, this rules out universality. Likewise, if $\sigma = p(z, \overline{z})$ is a polynomial, it is straightforward to see that each $\Psi \in \mathcal{NN}^1_{\sigma,L}$ is of the form $\Psi(z) = q_\Psi(z, \overline{z})$ for a complex polynomial q_Ψ of two variables and of degree $\deg q_\Psi \leq N$, with N only depending on the depth L and the degree of p; this again rules out universality.
- If σ is discontinuous but satisfies properties *a)* or *b)*, the theorem is not sufficient to decide whether $\mathcal{NN}^d_{\sigma,L}$ is universal or not. This is *not* a proof artifact; in fact, [12, Example 4.13] provides an example of such an activation function for which $\mathcal{NN}^d_{\sigma,L}$ is in fact universal for $L \geq 2$. Such activation functions are somewhat artificial, however: They are discontinuous but agree almost everywhere with a continuous function; hence, it would be natural to replace σ with its "continuous version."

The proofs of Theorems 2 and 3 are based on generalizing the arguments in [8] to the complex-valued case via the *Wirtinger calculus* and *Weyl's lemma.* For instance, using the identity $\Delta = 4\,\partial\overline{\partial}$ and elementary properties of the Wirtinger derivatives ∂ and $\overline{\partial}$, one can show that if $\Delta^m \sigma \equiv 0$, then also $\Delta^m \Psi \equiv 0$ for all $\Psi \in \mathcal{NN}^1_{\sigma,1}$. By Weyl's lemma, this extends to locally uniform limits. Thus, any $f : \mathbb{C} \to \mathbb{C}$ *not* satisfying $\Delta^m f \equiv 0$ cannot be approximated by elements of $\mathcal{NN}^1_{\sigma,1}$.

The universality of complex-valued neural networks has already been studied in the literature to some extent. However, the paper [12] is the first to provide a comprehensive characterization of the class of "universal activation functions." As a brief account of the literature, we mention that in [1] it was shown that the activation function σ defined by $\sigma(z) = 1/\big(1 + \exp(-\operatorname{Re}(z))\big) + i/\big(1 + \exp(-\operatorname{Im}(z))\big)$ gives rise to a universal class of shallow complex-valued networks. Moreover, the papers [6, 5] claim to prove universality of complex-valued networks for a variety of activation functions. However, some of these activation functions are in fact holomorphic, which shows that the arguments in [6, 5] cannot possibly be correct.

References

[1] P. Arena, L. Fortuna, G. Muscato, and M.G. Xibilia, *Neural networks in multidimensional domains: Fundamentals and new trends in modelling and control.* Springer-Verlag, 1998.

[2] G. Cybenko, *Approximation by superpositions of a sigmoidal function*, Math. Control Signal Systems **2** (1989), 303–314.

[3] K. Hornik, *Approximation capabilities of multilayer feedforward networks*, Neural Networks **4** (1991), 251–257.

[4] K. Hornik, M. Stinchcombe, and H. White, *Multilayer feedforward networks are universal approximators*, Neural Networks **2** (1989), 359–366.

[5] G.-B. Huang, M.-B. Li, L. Chen, and C.-K. Siew, *Incremental extreme learning machine with fully complex hidden nodes*, Neurocomputing **71** (2008), 576–583.

[6] T. Kim and T. Adalı, *Approximation by Fully Complex Multilayer Perceptrons*, Neural Computation **15** (2003), 1641–1666.

[7] Y. LeCun, Y. Bengio, and G. Hinton, *Deep Learning*, Nature **521** (2015), 436–444.

[8] M. Leshno, V. Y. Lin, A. Pinkus, and S. Schocken, *Multilayer feedforward networks with a nonpolynomial activation function can approximate any function*, Neural Networks **6** (1993), 861–867.

[9] H. Lin and S. Jegelka, *ResNet with one-neuron hidden layers is a universal approximator*, In Proceedings of the 32nd International Conference on Neural Information Processing Systems.

[10] C. Trabelsi, O. Bilaniuk, Y. Zhang, D. Serdyuk, S. Subramanian, J.F. Santos, S. Mehri, N. Rostamzadeh, Y. Bengio, and C.J. Pal, *Deep Complex Networks*, International Conference on Learning Representations (2018), URL: `openreview.net/forum?id=H1T2hmZAb`.

[11] P. Virtue, S. X. Yu, and M. Lustig, *Better than real: Complex-valued neural nets for MRI fingerprinting*, 2017 IEEE International Conference on Image Processing (2017), 3953–3957.

[12] F. Voigtlaender, *The universal approximation theorem for complex-valued neural networks*, arXiv preprints, arXiv:2012.03351.

[13] M. Wolter and A. Yao, *Complex Gated Recurrent Neural Networks*, NeurIPS (2018), 10557–10567.

[14] D.-X. Zhou, *Universality of deep convolutional neural networks*, Applied and Computational Harmonic Analysis **48** (2020), 787–794.

Supervised Learning for Maps between Banach Spaces

Nikola Kovachki

A general framework for data-driven approximation of input-output maps between infinite-dimensional spaces is developed. Motivated by the recent successes of neural networks, the proposed approach uses a combination of ideas from deep learning and model reduction. This combination results in a neural network approximation which, in principle, is defined on infinite-dimensional spaces and, in practice, is robust to the dimension of the finite-dimensional approximations of these spaces required for computation. For large classes of input-output maps, and suitably chosen probability measures on the inputs, convergence of the proposed approximation methodology is proved. Numerically, the effectiveness of the method is demonstrated on classes of parametric PDE problems with applications in reservoir modeling, the deformation of plastic materials, and the turbulent flow of fluids. Convergence and robustness of the approximation scheme with respect to the size of the discretization is established. The method is shown to be faster and more accurate than many existing algorithms in the literature.

References

[1] K. Bhattacharya, B. Hosseini, N. B. Kovachki, A. M. Stuart, *Model Reduction and Neural Networks for Parametric PDEs*, SMAI-JCM **7** (2021), 121–157.

[2] Z. Li, N. Kovachki, K. Azizzadenesheli, B. Liu, K. Bhattacharya, A. Stuart, A, Anandkumar, *Fourier Neural Operator for Parametric Partial Differential Equations*, In The International Conference on Learning Representations (ICLR 2021).

From linear to nonlinear n-widths: Optimality and Reduced Modelling

Albert Cohen

The concept of n-widths has been introduced by Kolmogorov as a way of measuring the size of compact sets in terms of their approximability by linear spaces. The n-width of a compact set K in a Banach space V is defined by

$$d_n(K) := \inf_{\dim(V_n)=n} \max_{u\in K} \min_{v\in V_n} \|u-v\|_V.$$

From a numerical perspective, it may be thought as a benchmark for the performance of algorithms based on linear approximation. In recent years, this concept has proved to be highly meaningful for the analysis of reduced modeling strategies in complex physical problems described by parametric PDE's.

On the one hand several results have demonstrated that certain families of parametrized PDE's have fast decaying n-width, in particular much faster than the decay that could be predicted from their standard spatial smoothness analysis in Sobolev spaces. Central to this state of affair is the fact that holomorphic maps preserve the rate of decay of n-widths : it was proved in [2] that if F is a map between Banach spaces V_1 and V_2 that is holomorphic in a neigbourhood of a compact set $K_1 \subset V_1$, then with $K_2 = F(K_1)$ one has

$$\sup_{n>0} n^s d_n(K_1)_{V_1} < \infty \implies \sup_{n>0} n^t d_n(K_2)_{V_2} < \infty,$$

for $t < s-1$. The loss of 1 in the rate may be an artifact of the proof and removing it is an open problem. This result can be applied to elliptic and parabolic PDEs, when F is the mapping that takes the diffusion function to the solution.

On the other hand, while the optimal n-width spaces are out of reach, they can be emulated by a greedy algorithm that iteratively selects $u^1, \dots, u^n$ from the compact set K and define a reduced basis space $V_n = \text{span}\{u^1, \dots, u^n\}$. Then the accuracy $\sigma_n(K)_V := \max_{u\in K} \min_{v\in V_n} \|u-v\|_V$ achieved by these spaces was proved in [1, 5] to be rate optimal in the sense that

$$\sup_{n>0} n^s d_n(K)_V < \infty \implies \sup_{n>0} n^s \sigma_n(K)_V < \infty,$$

and similar results hold for exponential rates.

It is however well-known that linear approximation methods perform poorly for certain classes of functions that exhibit singularities at arbitrarily points. Such classes typically occur when considering hyperbolic transport PDE's with shock positions that depend on the various parameters (flux, initial condition..). The linear n-widths of such classes decay poorly. This motivates for the use of nonlinear approximation strategies such as adaptive mesh refinement, best n-term approximation in a basis or dictionnary, rational fractions, neural networks.

Several attempts have been made to derive notions of n-widths that describe the optimal performance of nonlinear approximation methods. In particular, manifold widths are been defined in [6] as

$$\delta_n(K)_V := \inf_{E,D} \max_{u\in K} \|u - D(E(u))\|_V,$$

where the infimum is taken over all continuous encoding maps $E : V \to \mathbb{R}^n$ and decoding maps $D : \mathbb{R}^n \to V$. The continuity requirement is critical in order to avoid space filling manifolds which would make $\delta_n(K)_V$ a trivial quantity.

Manifold widths give a satisfactory description of nonlinear approximability for classical smoothness classes such as Besov spaces in the sense that they indeed reflect the rate of approximation achieved by methods such as adaptive mesh refinement or best n-term wavelet approximation. On the other hand, they do not reflect the capability of stable algorithms since continuity is a very weak assumption. This has motivated the introduction of stable nonlinear widths $\delta_{n,L}(K)_V$ that are defined in [3] similarly as $\delta_n(K)_V$, with the additional prescription that E and D should be Lipschitz continuous which constant bounded by L independently of n.

One main result is that when V is a Hilbert space, stable nonlinear widths are strongly tied to the Kolmogorov entropy numbers $\varepsilon_n(K)_V$ that are defined as the smallest $\varepsilon > 0$ such that K can be covered by 2^n balls of radius ε. Indeed, one has

$$\sup_{n>0} n^s \delta_{n,L}(K)_V < \infty \iff \sup_{n>0} n^s \varepsilon_n(K)_V < \infty.$$

This tight connection allows one to prove that the stable nonlinear n-widths of solution manifolds related to transport equations with shocks have fast decay, in contrast to their linear n-widths. A more general consequence is that if F is a map between a Banach space V_1 and a Hilbert space V_2 that is Lipschitz continuous over of a compact set $K_1 \subset V_1$, then with $K_2 = F(K_1)$ one has

$$\sup_{n>0} n^s \delta_{n,L}(K_1)_{V_1} < \infty \implies \sup_{n>0} n^s \delta_{n,L}(K_2)_{V_2} < \infty.$$

In this sense, stable nonlinear widths reflect the potential gain of nonlinear approximation methods over their linear counterpart for certain classes of PDEs.

On the other hand, recent results from [7, 4] showing that sampling numbers can be controlled by linear n-widths, and therefore that best linear approximation performance can be recovered from point value evaluation, do not seem to carry over to nonlinear widths.

References

[1] P. Binev, A. Cohen, W. Dahmen, R. DeVore, G. Petrova, and P. Wojtaszczyk *Convergence Rates for Greedy Algorithms in Reduced Basis Methods* (2011) SIAM J. Math. Anal. **43**,1457–1472.

[2] A. Cohen, R. DeVore, *Kolmogorov widths under holomorphic mappings*, IMA Journal of Numerical Analysis (2016) **36**, 1–12.

[3] A. Cohen, R. DeVore, G. Petrova, and P. Wojtaszczyk, *Optimal stable nonlinear approximation*, to appear in Foundation of Computational Mathematics (2021).

[4] A. Cohen and M. Dolbeault, *Optimal pointwise sampling for L^2 approximation*, to appear in Journal of Complexity (2021).

[5] R. DeVore, G. Petrova, and P. Wojtaszczyk *Greedy algorithms for reduced bases in Banach spaces*, J. of FoCM (2013) **37**, 455–466.

[6] R. DeVore, R. Howard, and C. Micchelli, *Optimal nonlinear approximation*, Manuscripta Mathematica (1989) **63**, 469–478.

[7] V. N. Temlyakov, *On optimal recovery in L^2*, preprint (2020).

Reporter: Anna Weller

Participants

Dr. Ben Adcock
Department of Mathematics and
Statistics
Simon Fraser University
Burnaby BC V5A 1S6
CANADA

Prof. Dr. Markus Bachmayr
Institut für Mathematik
Johannes-Gutenberg Universität Mainz
Staudingerweg 9
55128 Mainz
GERMANY

Prof. Dr. Richard Baraniuk
Dept. of Electrical & Computer
Engineer.
Rice University
6100 South Main Street
Houston, TX 77251-1892
UNITED STATES

Prof. Dr. Peter Binev
Department of Mathematics
University of South Carolina
Columbia, SC 29208
UNITED STATES

Prof. Dr. Andrea Bonito
Department of Mathematics
Texas A&M University
3368 TAMU
College Station TX, 77843-3368
UNITED STATES

Dr. Claire Boyer
Laboratoire de Probabilités, Statistique
et Modélisation (LPSM), Case 247
Sorbonne Université
Campus Pierre et Marie Curie
4, Place Jussieu
75252 Paris Cedex 05
FRANCE

Prof. Dr. Claudio Canuto
Dipartimento di Scienze Matematiche
Politecnico di Torino
Corso Duca degli Abruzzi, 24
10129 Torino
ITALY

Prof. Dr. Albert Cohen
Laboratoire Jacques-Louis Lions
Sorbonne Université
4, Place Jussieu
75005 Paris Cedex
FRANCE

Prof. Dr. Wolfgang Dahmen
Department of Mathematics
University of South Carolina
1523 Greene Street
Columbia, SC 29208
UNITED STATES

Prof. Dr. Ronald A. DeVore
Department of Mathematics
Texas A & M University
College Station, TX 77843-3368
UNITED STATES

Matthieu Dolbeault
Laboratoire Jacques-Louis Lions
Université Pierre et Marie Curie
4, Place Jussieu
75005 Paris
FRANCE

Prof. Dr. Simon Foucart
Department of Mathematics
Texas A & M University
College Station, TX 77843-3368
UNITED STATES

Prof. Dr. Lars Grasedyck
Institut für Geometrie und
Praktische Mathematik
RWTH Aachen
Templergraben 55
52062 Aachen
GERMANY

Prof. Dr. Philipp Grohs
Fakultät für Mathematik
Universität Wien
Kolingasse 14-16
1090 Wien
AUSTRIA

Prof. Dr. László Györfi
Department of Computer Science and
Information Theory
Budapest University of Technology
and Economics
Stoczek u. 2
1521 Budapest
HUNGARY

Prof. Dr. Helmut Harbrecht
Departement Mathematik und
Informatik
Universität Basel
Spiegelgasse 1
4051 Basel
SWITZERLAND

Laslo Hunhold
Department Mathematik/Informatik
Abteilung Mathematik
Universität zu Köln
Weyertal 86-90
50931 Köln
GERMANY

Dr. Yoshihito Kazashi
Institute of Mathematics, École
polytechnique fédérale de Lausanne
EPFL SB MATH CSQI
MA B2 434 (Bâtiment MA)
Station 8
1015 Lausanne
SWITZERLAND

Prof. Dr. Gerard Kerkyacharian
Laboratoire de Probabilites, Tour 56
Université P. et M. Curie
4, Place Jussieu
75252 Paris Cedex 05
FRANCE

Nikola B. Kovachki
California Institute of Technology
MC 305-16
1200 E. California Boulevard
Pasadena CA 91125
UNITED STATES

Prof. Dr. Angela Kunoth
Department Mathematik/Informatik
Universität zu Köln
Weyertal 86-90
50931 Köln
GERMANY

Prof. Dr. Gitta Kutyniok
Mathematisches Institut
Ludwig-Maximilians-Universität
München
Theresienstraße 39
80333 München
GERMANY

Prof. Dr. Tony Lelievre
CERMICS - ENPC
Cite Descartes, Champs-sur-Marne
6 et 8 Avenue Blaise Pascal
77455 Marne-la-Vallée Cedex 2
FRANCE

Prof. Dr. Wuchen Li
Department of Mathematics
University of South Carolina
Columbia, SC 29208
UNITED STATES

Dr. Damiano Lombardi
INRIA Paris
2, Rue Simone Iff
75012 Paris Cedex
FRANCE

Prof. Dr. Siddhartha Mishra
Departement Mathematik
ETH-Zentrum
Rämistrasse 101
8092 Zürich
SWITZERLAND

Prof. Ricardo H. Nochetto
Department of Mathematics
Institute for Physical Science and Technology
University of Maryland
College Park MD 20742-2431
UNITED STATES

Prof. Dr. Anthony Nouy
Ecole Centrale de Nantes, GeM
1, rue de la Noe
P.O. Box 92101
44321 Nantes Cedex 3
FRANCE

Prof. Dr. Peter Oswald
Institut für Numerische Simulation
Universität Bonn
Friedrich-Hirzebruch-Allee 7
53115 Bonn
GERMANY

Dr. Josiah Park
Department of Mathematics
Texas A & M University
College Station, TX 77843-3368
UNITED STATES

Dr. Philipp Christian Petersen
Fakultät für Mathematik
Universität Wien
Kolingasse 14-16
1090 Wien
AUSTRIA

Prof. Dr. Guergana Petrova
Department of Mathematics
Texas A & M University
College Station, TX 77843-3368
UNITED STATES

Dr. Pencho Petrushev
Department of Mathematics
University of South Carolina
1523 Greene Street
Columbia SC 29208
UNITED STATES

Prof. Dr. Dominique Picard
LPSM
21 rue des plantes
75014 Paris Cedex
FRANCE

Prof. Dr. Holger Rauhut
LST Mathematik der
Informationsverarbeitung
RWTH Aachen
Pontdriesch 10
52062 Aachen
GERMANY

Prof. Dr. Lars Ruthotto
Department of Mathematics
Emory University
400 Dowman Drive
Atlanta GA 30322
UNITED STATES

Prof. Dr. Robert Scheichl
Institut für Angewandte Mathematik
Universität Heidelberg
Im Neuenheimer Feld 205
69120 Heidelberg
GERMANY

Prof. Dr. Johannes Schmidt-Hieber
Department of Applied Mathematics
University of Twente
P.O.Box 217
7500 AE Enschede
NETHERLANDS

Prof. Dr. Reinhold Schneider
Fakultät II - Institut für Mathematik
Technische Universität Berlin
Sekr. MA 5 - 3
Straße des 17. Juni 136
10623 Berlin
GERMANY

Agustin Somacal
Sorbonne Université
Laboratoire d'Informatique de Paris 6
4 Place Jussie
P.O. Box 169
75252 Paris
FRANCE

Prof. Dr. Rob P. Stevenson
Korteweg de Vries Instituut
Universiteit van Amsterdam
P.O. Box 94248
1090 GE Amsterdam
NETHERLANDS

Prof. Dr. Endre Süli
Mathematical Institute
University of Oxford
Radcliffe Observatory Quarter
Woodstock Road
Oxford OX2 6GG
UNITED KINGDOM

Prof. Dr. Karsten Urban
Institut für Numerische Mathematik
Universität Ulm
Helmholtzstrasse 20
89081 Ulm
GERMANY

Dr. Felix Voigtlaender
Zentrum Mathematik
Technische Universität München
Boltzmannstrasse 3
85748 Garching bei München
GERMANY

Anna Weller
Department Mathematik/Informatik
Abteilung Mathematik
Universität zu Köln
Weyertal 86 - 90
50931 Köln
GERMANY

Dr. Gerrit Welper
Department of Mathematics
University of Central Florida
4393 Andromeda Loop N
Orlando FL 32816
UNITED STATES

Prof. Dr. Przemek Wojtaszczyk
Institute of Mathematics, Polish
Academy of Sciences
Śniadeckich 8
00-656 Warszawa
POLAND

Prof. Dr. Jinchao Xu Jinchao Xu
Department of Mathematics
Pennsylvania State University
University Park PA 16802
UNITED STATES

Tolunay Yilmaz
Department Mathematik/Informatik
Universität zu Köln
Weyertal 86-90
50931 Köln
GERMANY