Mathematisches Forschungsinstitut Oberwolfach

Report No. 33/2016

DOI: 10.4171/OWR/2016/33

Learning Theory and Approximation

Organised by
Andreas Christmann, Bayreuth
Kurt Jetter, Stuttgart
Steve Smale, Hong Kong and Berkeley
Ding-Xuan Zhou, Hong Kong

3 July – 9 July 2016

Abstract. The main goal of this workshop – the third one of this type at the MFO – has been to blend mathematical results from statistical learning theory and approximation theory to strengthen both disciplines and use synergistic effects to work on current research questions. Learning theory aims at modeling unknown function relations and data structures from samples in an automatic manner. Approximation theory is naturally used for the advancement and closely connected to the further development of learning theory, in particular for the exploration of new useful algorithms, and for the theoretical understanding of existing methods. Conversely, the study of learning theory also gives rise to interesting theoretical problems for approximation theory such as the approximation and sparse representation of functions or the construction of rich kernel reproducing Hilbert spaces on general metric spaces. This workshop has concentrated on the following recent topics: Pitchfork bifurcation of dynamical systems arising from mathematical foundations of cell development; regularized kernel based learning in the Big Data situation; deep learning; convergence rates of learning and online learning algorithms; numerical refinement algorithms to learning; statistical robustness of regularized kernel based learning.

Mathematics Subject Classification (2010): 68Q32, 41A35, 41A63, 62Gxx.

Introduction by the Organisers

The workshop "Learning Theory and Approximation", organized by Andreas Christmann (Bayreuth), Kurt Jetter (Stuttgart-Hohenheim), Steve Smale (Hong Kong and Berkeley), and Ding-Xuan Zhou (Hong Kong), was held July 3–9, 2016.

This half-size workshop was well attended, with 26 participants from Asia, Europe and North America. It provided an excellent platform for fruitful interactions among researchers from statistical learning theory and approximation theory.

Learning theory aims at modeling unknown function relations and data structures from samples in an automatic manner. It started with some topics in statistics such as pattern recognition, nonparametric estimation, support vector machines, and statistical learning theory. More connections and applications to other fields have been found within the last decade: computational biology, data mining, computational topology, optimization theory, ranking methods, survival statistics, and many others.

Already the first talks on Monday showed the broad coverage of this workshop and opened discussions on the interplay between approximation theory, statistical learning theory, dynamical systems, and computational biology. Steve Smale's talk on pitchfork bifurcation of dynamical systems arose from his mathematical foundations to understand cell development in multi cellulars organisms. A toggle switch of two gene networks was modelled by a special system of differential equations having second order terms which is in contrast to a classical model proposed in the literature. In this talk it was proven that generically this new dynamical system undergoes pitchfork bifurcation. Tomaso Poggio's talk compared recent deep learning algorithms and more classical shallow networks. His talk demonstrated that deep learning methods can often outperform shallow networks if the underlying unknown function is of compositional type whereas no substantial improvement seems to be possible in the general case because good shallow networks are universally consistent. Recent oracle inequalities and learning rates for binary classification algorithms using adaptive partitioning were presented by Peter Binev. The considered new approximation classes are much richer than Besov spaces.

The Monday afternoon session was on approximation theory. The talk by Gerlind Plonka-Hoch proposed a sparse approximation of structured signals by the modified Prony method and an explicit algorithm for sparse approximation of exponential sums was presented. Holger Wendland's talk was on recent results of multiscale radial basis functions. These meshfree methods occur in many disciplines including scattered data approximation, statistical machine learning, engineering, and computer graphics. The results covered matrix-valued kernels, which lead to divergence-free approximation spaces and their multiscale extensions. Interpolation and quasi-interpolation with multiquadrics were the topic of Martin Buhmann's talk. He showed that good error bounds are achievable, even when the usual additonal constant is not present in the ansatz. To manage this modification, he had to employ native spaces of Pontryagin type.

On Tuesday there were talks on learning theory and on approximation theory. Bernhard Schölkopf emphasized in his talk that statistical correlation does not necessarily imply causality and demonstrated how causality can be checked using statistical learning. Johan Suykens showed new extensions of learning with

primal and dual model representations, in particular multilevel hierarchical kernel spectral clustering for large scale networks and deep learning using restricted kernel machines and conjugate feature duality. Alexandre Tsybakov considered structured high-dimensional least squares estimation and proved new minimax optimal results. Standard kernel approaches often fail for the Big Data situation, because they do not scale well with the number of sample points. Therefore, Ding-Xuan Zhou gave detailed error analyses for broad classes of distributed learning algorithms based on a divide-and-conquer approach including least squares regularization schemes and spectral algorithms. Gabriele Steidl presented new results on iterative multiplicative filters for data labeling, which has been successfully applied, e.g, for image partioning and segmentation. She proposed an algorithm that can be seen as an iterative multiplicative filtering of a label assignment matrix. Philipp Kügler showed how action potential dynamics can be learnt for preclinical drug safety testing. These results are important for pharmaceutical companies.

The talks on Wednesday morning were on machine learning. Ingo Steinwart demonstrated in his talk on learning with hierarchical kernels, that a data-dependent weighted sum of Gaussian kernels has a reproducing kernel Hilbert space which is dense in $C(X)$ with respect to the supremum norm provided the input space X is a compact subset of $\mathbb{R}^d$. This talk was strongly related to Tomaso Poggio's talk on deep learning, but the point of view was clearly quite different. It became clear that more theoretical work is desirable for both approaches. Sayan Mukherjee proved the asymptotic consistency of maximum likelihood estimators for dynamical systems observed with noise. The proof involved ideas from both information theory and dynamical systems. Examples were shifts of finite type with Gibbs measures and Axiom A attractors with SRB measures. Andreas Christmann presented joint work with Ding-Xuan Zhou on robust pairwise learning with kernels. Examples of pairwise learning occur in ranking problems and minimum error entropy estimation. The robustness results covered a bounded influence function, upper bounds for the maxbias over neighborhoods of total variation and of gross-error neighborhoods, and qualitative robustness of the kernel estimators and their empirical bootstrap approximation.

The first talks on Thursday dealt with approximation theory and its relationship to machine learning. Compressive sensing considers the recovery of (approximately) sparse vectors (signals, images etc.) from incomplete linear measurements via efficient algorithms. An extension of this theory replaces the sparsity assumption by a low rank assumption of a matrix to be recovered. Holger Rauhut's talk investigated the sparse and low rank recovery via Mendelson's small ball method. Uniform convergence of stationary d-variate subdivision with a finite mask can be analyzed through left convergence of products of certain matrices from a finite alphabet of matrices constructed from the mask. In his talk on nonnegative subdivision, Kurt Jetter showed that uniform convergence of nonnegative subdivision is equivalent to the fact that each word from this alphabet is stochastic, indecomposable and aperiodic. Equivalently, each word of sufficient length must have the scrambling property, or even must have a strictly positive column. The latter two

properties refer to sign patterns of row stochastic matrices, and sufficient length means that the length is at most equal to the number of possible sign patterns of such matrices. Tomas Sauer's talk on the recovery of sparse exponential sums and sparse polynomials in several variables had strong connections to the talk by Gerlind Plonka-Hoch. After a deep analysis it became clear, that the computation of a special Prony ideal is crucial, because then the frequencies can be determined by eigenvalue methods analogous to Frobenius companion matrices and the coefficients by solving a special Vandermonde system. Stéphane Boucheron studied the problem of lossless universal source coding for stationary memoryless sources on a countably infinite alphabet and proposed an adaptive compression technique: a collection of so-called envelope classes is considered and both dictionary and pattern encoding are treated.

The talks on Thursday afternoon mainly dealt with machine learning and wavelets. Yiming Ying presented his recent work on online learning with pairwise loss functions. Pairwise learning differs from more traditional learning tasks like classification or regression, because (i) the objective function is usually defined over pairs of instances which is quadratic in the sample size, and (ii) pairwise learning involves statistically dependent pairs of data points, which is fundamentally different from the i.i.d. assumption in classification and regression. Yiming Ying showed that the algorithmic implementation and the theoretical analysis of his method is comparable to online algorithms in classification. This talk had obvious connections to the talk by Andreas Christmann on the robustness aspects of (non-online) pairwise learning. Dao-Hong Xiang presented her recent work on quantile regression with varying Gaussians, coefficient-based conditional quantile regression and learning with varying ϵ-insensitive pinball loss. The convergence of the randomized Kaczmarz algorithm in Hilbert spaces was investigated by Xin Guo. The convergence is a weak convergence with a polynomial rate. Weak convergence is widely used in learning theory because it well corresponds to the strong convergence in the L_2 norm sense which is usually good enough for applications.

Friday was reserved for approximation theory. Joachim Stöckler combined methods of real algebraic geometry, linear system theory and harmonic analysis for the construction and parameterization of classes of tight wavelet frames. Maria Charina investigated the construction of orthogonal multi-wavelets. She showed that there is no much conceptual difference between wavelet ($n = m = 1$) and multi-wavelet constructions and provided their complete and unifying characterization. This characterization is based on classical results from system theory. The link between wavelet and multi-wavelet constructions and system theory is offered by the so-called Unitary Extension Principle. Karlheinz Gröchenig investigated the question how many samples of a function f are necessary to completely recover this function. By generalizing the Beurling concept of lower and upper density, he derived quite general theorems for the study of sets of sampling, and of sets of interpolation, in the setting of reproducing kernel Hilbert spaces. In this way, universal density theorems are produced which include the results from the literature as special cases. Elena Berdysheva's talk on Durrmeyer type operators

with respect to an arbitrary measure was motivated by earlier work by Zhou and Jetter (2006) in the context of support vector machine classifiers with polynomial kernels. This operator is a special compact self-adjoint integral operator and its kernel is a Mercer kernel. In her talk, she dealt with various convergence properties of this operator.

Acknowledgement: The organizers acknowledge the friendly atmosphere provided by the Oberwolfach institute, and would like to express their thanks to the entire staff. The MFO and the workshop organizers would like to thank the National Science Foundation for supporting the participation of junior researchers in the workshop by the grant DMS-1049268, "US Junior Oberwolfach Fellows".

Workshop: Learning Theory and Approximation

Table of Contents

Abstracts

Pitchfork Bifurcation

Steve Smale

Development of a single cell into a multicellular organism involves a remarkable integration of gene expression, molecular signaling, and environmental cues. This talk is about pitchfork bifurcation of dynamical systems arising from our study of mathematical foundations of cell development. This is joint work with I. Rajapakse and is related to our previous work [1].

A classical example of pitchfork bifurcation (e.g. [2]) is from the following system of ordinary differential equations

$$\frac{dx}{dt} = \mu x - x^3, \qquad x \in \mathbb{R} \tag{1}$$

with a parameter $\mu \in \mathbb{R}$. This system has an equilibrium $x_0 = 0$ for all μ. This equilibrium is stable for $\mu < 0$ and unstable for $\mu > 0$. For $\mu > 0$, there are two extra equilibria, $x_{1,2} = \pm\sqrt{\mu}$, branching from the origin which are stable. This bifurcation is called pitchfork bifurcation.

A toggle switch of two gene networks described by

$$\begin{aligned} \frac{dx}{dt} &= \frac{2}{1+y^m} - x, \\ \frac{dy}{dt} &= \frac{2}{1+x^m} - y \end{aligned} \tag{2}$$

was designed and constructed in [3]. It was used to predict conditions for bistability and pitchfork bifurcation, which was proved in [4].

By taking Taylor expansions of order two at $m = 2, x = y = 1$ for the Hill functions in (2), we derive the following dynamical system

$$\begin{aligned} \frac{dx}{dt} &= y^2 - my - x, \\ \frac{dy}{dt} &= x^2 - mx - y. \end{aligned}$$

This system has second order terms, which is different from the classical setting (1). We prove that generically this dynamical system undergoes pitchfork bifurcation.

References

[1] I. Rajapakse and S. Smale, *Mathematics of the genome*, preprint, 2016.

[2] J. Guckenheimer and P. Holmes, *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, Springer, New York, 2002.

[3] T. S. Gardner, C. R. Cantor, and J. J. Collins, *Construction of a genetic toggle switch in Escherichia coli*, Nature, **403** (2000), 339–342.

[4] S. P. Ellner and J. Guckenheimer, *Dynamics Models in Biology*, Princeton University Press, Princeton, New Jersey, 2006.

Deep vs. shallow networks: An approximation theory perspective

Tomaso Poggio
(joint work with H. N. Mhaskar)

Summary
We describe recent results on hierarchical architectures for learning from examples, that may formally explain the conditions under which Deep Convolutional Neural Networks perform much better in function approximation problems than shallow, one-hidden layer architectures.

Introduction
Deep Neural Networks especially of the convolutional type (DCNNs) have started a revolution in the field of artificial intelligence and machine learning, triggering a large number of commercial ventures and practical applications. Most deep learning references these days start with Hinton's backpropagation and with Lecun's convolutional networks (see for a nice review [4]). Of course, multilayer convolutional networks have been around at least as far back as the optical processing era of the 70s. Fukushima's Neocognitron [2] was a convolutional neural network that was trained to recognize characters. The HMAX model of visual cortex [8] was described as a series of AND and OR layers to represent hierarchies of disjunctions of conjunctions.

Two of the basic theoretical questions about Deep Convolutional Neural Networks (DCNNs) are:

- which classes of functions can they approximate well?
- why is stochastic gradient descent (SGD) so unreasonably efficient?

In this contribution we describe a theoretical framework that we have introduced very recently to address the first question [6]. The theoretical results include answers to why and when deep networks are better than shallow by using the idealized model of a deep network as a directed acyclic graph (DAG), which we have shown to capture the properties a range of convolutional architectures recently used, such as the very deep convolutional networks of the ResNet type [3]. For compositional functions conforming to a DAG structure with a small maximal indegree of the nodes, such as a binary tree structure, one can bypass the curse of dimensionality with the help of the blessings of compositionality (cf. [1] for a motivation for this terminology). We demonstrate this fact using three examples: traditional sigmoidal networks, the ReLU networks commonly used in DCNN's, and Gaussian networks.

Compositional functions
Let us illustrate the advantage of approximating a compositional function using deep networks corresponding to the compositional structure rather than a shallow network that does not take into account this structure.

In the sequel, for any integer $q \geq 1$, $\mathbf{x} = (x_1, \cdots, x_q) \in \mathbb{R}^q$, $|\mathbf{x}|$ denotes the Euclidean ℓ^2 norm of $\mathbf{x}$, and $\mathbf{x} \cdot \mathbf{y}$ denotes the usual inner product between $\mathbf{x}, \mathbf{y} \in \mathbb{R}^q$. In general, we will not complicate the notation by mentioning the dependence on the dimension in these notations unless this might lead to confusion.

Let $I^q = [-1,1]^q$, $\mathbf{X} = C(I^q)$ be the space of all continuous functions on I^q, with $\|f\| = \max_{\mathbf{x}\in I^q} |f(\mathbf{x})|$. Let $\mathcal{S}_n$ denote the class of all shallow networks with n units of the form

$$\mathbf{x} \mapsto \sum_{k=1}^{n} a_k \sigma(\mathbf{w}_k \cdot \mathbf{x} + b_k),$$

where $\mathbf{w}_k \in \mathbb{R}^q$, $b_k, a_k \in \mathbb{R}$. The number of trainable parameters here is $(q+2)n \sim n$. Let $r \geq 1$ be an integer, and $W_{r,q}^{\mathrm{NN}}$ be the set of all functions with continuous partial derivatives of orders up to r such that $\|f\| + \sum_{1\leq|\mathbf{k}|_1\leq r} \|D^{\mathbf{k}} f\| \leq 1$, where $D^{\mathbf{k}}$ denotes the partial derivative indicated by the multi–integer $\mathbf{k} \geq 1$, and $|\mathbf{k}|_1$ is the sum of the components of $\mathbf{k}$.

For explaining our ideas for the deep network, we consider compositional functions conforming to a binary tree. For example, we consider functions of the form (cf. Figure 2)

(1)
$$f(x_1, \cdots, x_8) = h_3(h_{21}(h_{11}(x_1, x_2), h_{12}(x_3, x_4)), h_{22}(h_{13}(x_5, x_6), h_{14}(x_7, x_8))).$$

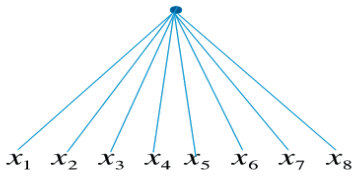

FIGURE 1. *A shallow universal network in 8 variables and N units which can approximate a generic function* $f(x_1, \cdots, x_8)$. *The top node consists of n units and computes the ridge function* $\sum_{i=1}^{n} a_i \sigma(\langle \mathbf{v}_i, \mathbf{x} + t_i)\rangle$, *with* $\mathbf{v}_i, \mathbf{x} \in \mathbb{R}^2$, $a_i, t_i \in \mathbb{R}$.

For the hierarchical binary tree network, the spaces analogous to $W_{r,q}^{\mathrm{NN}}$ are $W_{H,r,2}^{\mathrm{NN}}$, defined to be the class of all functions f which have the same structure (e.g., (1)), where each of the constituent functions h is in $W_{r,2}^{\mathrm{NN}}$ (applied with only 2 variables). We define the corresponding class of deep networks $\mathcal{D}_n$ to be set of all functions with the same structure, where each of the constituent functions is in $\mathcal{S}_n$. We note that in the case when q is an integer power of 2, the number of parameters involved in an element of $\mathcal{D}_n$ – that is, weights and biases, in a node of the binary tree is $(q-1)(q+2)n$.

The following theorem (cf. [5]) estimates the degree of approximation for shallow and deep networks. We remark that the assumptions on σ in the theorem

below are not satisfied by the ReLU function $x \mapsto |x|$, but they are satisfied by smoothing the function in an arbitrarily small interval around the origin.

Theorem 1. *Let $\sigma : \mathbb{R} \to \mathbb{R}$ be infinitely differentiable, and not a polynomial on any subinterval of $\mathbb{R}$.*
(a) *For $f \in W_{r,q}^{NN}$*

$$\mathsf{dist}(f, \mathcal{S}_n) = \emptyset(n^{-r/q}). \tag{2}$$

(b) *For $f \in W_{H,r,2}^{NN}$*

$$\mathsf{dist}(f, \mathcal{D}_n) = \mathcal{O}(n^{-r/2}). \tag{3}$$

Proof. Theorem 1(a) was proved by [5]. To prove Theorem 1(b), we observe that each of the constituent functions being in $W_{r,2}^{\mathrm{NN}}$, (2) applied with $q = 2$ implies that each of these functions can be approximated from $\mathcal{S}_n$ up to accuracy $n^{-r/2}$. Our assumption that $f \in W_{H,r,2}^{\mathrm{NN}}$ implies that each of these constituent functions is Lipschitz continuous. Hence, it is easy to deduce that, for example, if P, P_1, P_2 are approximations to the constituent functions h, h_1, h_2, respectively within an

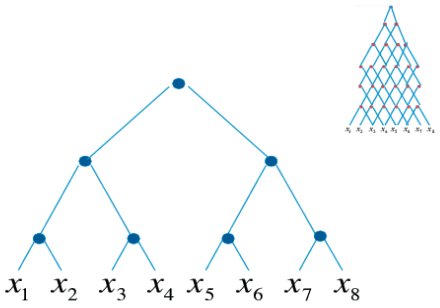

FIGURE 2. *A binary tree hierarchical network in 8 variables, which approximates well functions of the form (1). Each of the nodes consists of n units and computes the ridge function $\sum_{i=1}^{n} a_i \sigma(\langle \mathbf{v}_i, \mathbf{x} + t_i \rangle)$, with $\mathbf{v}_i, \mathbf{x} \in \mathbb{R}^2$, $a_i, t_i \in \mathbb{R}$. Similar to the shallow network such a hierarchical network can approximate any continuous function; the text proves how it approximates compositional functions better than a shallow network. Shift invariance may additionally hold implying that the weights in each layer are the same. The inset at the top right shows a network similar to ResNets: our results on binary trees apply to this case as well with obvious changes in the constants.*

accuracy of ϵ, then

$$\begin{aligned}\|h(h_1,h_2)-P(P_1,P_2)\| &\leq \|h(h_1,h_2)-h(P_1,P_2)\|+\|h(P_1,P_2)-P(P_1,P_2)\| \\ &\leq c\{\|h_1-P_1\|+\|h_2-P_2\|+\|h-P\|\}\leq 3c\epsilon,\end{aligned}$$

for some constant $c>0$ independent of ϵ. This leads to (3). □

We have extended the basic theorem to networks with ReLU and with Gaussian activation functions. We have also extended the result to general DAG functions that is functions defined on a directed acyclic graph (DAG)

As we mentioned in previous papers [7, 6] this definition, and in fact most of the previous results, can be specialized to the class of Boolean functions which map the Boolean cube into reals, yielding a number of known and new results. This application will be described in a forthcoming paper.

References

[1] D. L. Donoho et al. *High-dimensional data analysis: The curses and blessings of dimensionality*, AMS Math Challenges Lecture (2000), 1–32.

[2] K. Fukushima. *Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position*, Biological Cybernetics, **36(4)**, (1980), 193–202.

[3] K. He, X. Zhang, S. Ren, and J. Sun. *Deep residual learning for image recognition*, arXiv preprint arXiv:1512.03385v1 [cs.CV] 10 Dec 2015 (2015).

[4] Y. LeCun, Y. Bengio, and G. Hinton. *Deep learning*, Nature, **521(7553)**, (2015), 436–444.

[5] H. N. Mhaskar. *Neural networks for optimal approximation of smooth and analytic functions*, Neural Computation, **8(1)**, (1996), 164–177.

[6] H. N. Mhaskar, Q. Liao, and T. Poggio. *Learning real and boolean functions: When is deep better than shallow*, arXiv preprint arXiv:1603.00988, also Center for Brains, Minds and Machines (CBMM) Memo No. 45, (2016).

[7] T. Poggio, F. Anselmi, and L. Rosasco. *I-theory on depth vs width: hierarchical function composition*, CBMM memo 041, (2015).

[8] M. Riesenhuber and T. Poggio. *Hierarchical models of object recognition in cortex*, Nature Neuroscience, **2(11)** (1999), 1019–1025.

On classification algorithms using adaptive partitioning

Peter Binev

(joint work with Albert Cohen, Wolfgang Dahmen, Ronald DeVore)

We consider algorithms based on adaptive partitioning for classification of randomly drawn data. The setup for binary classification is the following. Let $X\in\mathbb{R}^d$, $Y=\{-1,1\}$, $Z=X\times Y$, and assume that $\rho=\rho_X\cdot\rho(y|x)$ is the probability measure on Z according to which the data is drawn. Denoting by $p(x)$ the probability that $y=1$ given x, we define the regression function $\eta(x):=\mathbb{E}(y|x)=2p(x)-1$ as the expectation of y given x. Any ρ_X-measurable set $\Omega\subset X$ can be considered a *classifier* assuming that it predicts $y=1$ for all $x\in\Omega$ and $y=-1$ for all $x\in\Omega^c:=X\setminus\Omega$. The probability of misclassification

$$R(\Omega):=\int_\Omega 1-p(x)d\rho_x+\int_{\Omega^c}p(x)d\rho_x$$

by a set Ω is called *risk*. The *Bayes classifier* $\Omega^* := \{x : \eta(x) \geq 0\}$ minimizes this risk and usually the performance of a classifier Ω is measured by the *excess risk*

$$R(\Omega) - R(\Omega^*) = \int_{\Omega \triangle \Omega^*} |\eta(x)| d\rho_X \ ,$$

where $A \triangle B := (A \setminus B) \cup (B \setminus A)$ is the symmetric difference of two sets. If Ω has to be chosen from a family of sets, then it is easy to see that the best possible choice would be the set Ω that maximizes the quantity $\eta_\Omega := \int_\Omega \eta(x) d\rho_X$.

A classification algorithm finds a classifier $\hat{\Omega}(\mathbf{z})$ based on given data $\mathbf{z} = (z_i)_{i=1}^n$ of points $z_i = (x_i, y_i) \in Z$ drawn independently according to ρ. Our approach to building such algorithms is to approximate directly the Bayes set $\Omega*$ using adaptively generated partitions of X. This type of algorithms are usually categorized as *set classifiers* in contrast to the *plug-in classifiers* which are based on the estimation of the regression function η. In both cases the performance of the algorithms is judged by how fast the excess risk decays when the sample size n grows. The derivation of estimates about this behavior are usually based on properties of the measure ρ quantified via assumptions on its behavior near the boundary of the set Ω^* (a margin condition) and the smoothness of the regression function η. A typical margin condition (see [4, 3]) is the requirement (also known as Tsybakov condition) that for some $\alpha \geq 0$ there exists a constant C_α such that

$$\rho_X\{x \in X : |\eta(x)| \leq t\} \leq C_\alpha t \,, \qquad 0 < t \leq 1 \,. \tag{1}$$

The smoothness conditions are often expressed via approximation classes that are linked to a nonlinear approximation process. In the case of set estimators, one can consider a nested sequence $(\mathcal{S}_m)_{m\geq 1}$ of families of subsets of X, where m represents the complexity of the family $\mathcal{S}_m$. The approximation error $a_m(\rho)$ is defined via how well the Bayes classifier Ω^* is approximated by the family $\mathcal{S}_m$

$$a_m(\rho) := \inf_{S \in \mathcal{S}_m} R(S) - R(\Omega^*) \,.$$

Using that this quantity is monotone with m, we define the approximation class $\mathcal{A}^s = \mathcal{A}^s\left((\mathcal{S}_m)_{m\geq 1}\right)$ as the set of all probability measures ρ for which the following semi-norm is finite

$$|\rho|_{\mathcal{A}^s} := \sup_{m \geq 1} m^s a_m(\rho) \,.$$

These approximation classes depend on how the families $\mathcal{S}_m$ are defined. In [1] the building of $\mathcal{S}_m$ is based on dyadic subdivision in which each element of the current partition is subdivided into 2^d subsets. Starting from the set X itself, one can perform $k \leq m$ such subdivisions to get a partition, split each element of the partition by an arbitrary hyperplane, and then choose any collection of the resulting subsets to form one set of the family $\mathcal{S}_m$. While the number of elements of $\mathcal{S}_m$ is infinite, it is proven in [1] that its VC dimension is limited by m times a constant depending only on d.

The classification algorithm is based on a model selection procedure. For this purpose we split the draw $\mathbf{z}$ into two independent equal parts and then use the first one to find the sets $\bar{\Omega}_m \in \mathcal{S}_m$ for $m \geq 1$ that maximize the empirical counterpart

of $\eta_{\bar{\Omega}_m}$. Then find the index m^* which maximizes the empirical quantity $\eta_{\bar{\Omega}_m}$ for $m \geq 1$ but this time based on the second part of the draw and set $\hat{\Omega} = \bar{\Omega}_{m*}$. Note that to determine the hyperplane for the hyperplane split, it is computationally more efficient to use a local plug-in estimator.

The performance of the above algorithm is given by the following result (see [1], Theorem 6.3).

Theorem (i) For any $r > 0$, there is a constant $c > 0$ such that the following holds. If $\rho \in \mathcal{A}^s$, $s > 0$, and ρ satisfies the margin condition (1), then with probability greater than $1 - cn^{-r+1}$, we have

$$R(\hat{\Omega}(\mathbf{z})) - R(\Omega^*) \leq C \left(\frac{\log n}{n} \right)^{\frac{(1+\alpha)s}{(2+\alpha)s+1+\alpha}}$$

with C depending only on d, r, $|\rho|_{\mathcal{A}^s}$ and the constant C_α in (1).

(ii) If $\eta \in B^\beta_\infty(L_p(X))$ with $0 < \beta \leq 2$ and $p > d/\beta$ and if ρ satisfies the margin condition (1), then with probability greater than $1 - cn^{-r+1}$, we have

$$R(\hat{\Omega}(\mathbf{z})) - R(\Omega^*) \leq C \left(\frac{\log n}{n} \right)^{\frac{(1+\alpha)\beta}{(2+\alpha)\beta+d}},$$

with C depending only on d, r, $|\eta|_{B^\beta_\infty(L_p(X))}$ and the constant C_α in (1).

Extensions of this result could come from improving the family $\mathcal{S}_m$. One possible way is the replacement of the hyperplane split at the end with a partitioning by a higher order polynomial surface. While this will increase the upper bound for β in (ii), the implementation of such a procedure is a demanding task and such a method does not seem practical. Another possibility is to revisit the adaptive partitioning procedure before the hyperplane splits. This procedure is equivalent to building a decision tree for the elements of the partition to be further subdivided. For large dimension d the practical solution is to consider only the nodes of the tree (aka elements of the partition) that contain data points, hence consideration of *occupancy* trees.

In going further, we consider binary trees instead of dyadic ones, replacing each dyadic split with a binary subtree that is a full binary tree with d levels, and then trim based on occupancy. This could give slight improvement of the computational efficiency but the major benefit comes from the better ways to handle sparsity when dealing with binary trees.

It is often the case in occupancy trees that a node has only one descendant for several generations, although going to the next generation is counted as a subdivision despite the fact that one of the elements is not occupied. We can therefore decrease significantly the complexity count m of the tree by introducing the notion of *sparse occupancy trees*, see [2], in which the sequences of nodes with a single descendant are collapsed to one node in the sparse occupancy tree.

In order to align the sparse occupancy with the probabilistic setup, we have to declare "unoccupied" sets S that have probability measure below some threshold $t > 0$. Note that t should be significantly larger than $\frac{1}{n}$, n being the sample size, in order to have an integer $\tau > 0$ such that if the number of sample points in a

given set S is less than τ, then with high probability $\int_S dr_X < t$, while if this number is at least τ, then $\int_S dr_X \geq \frac{t}{2}$ with high probability. Next, we create a partition by subdividing all the elements with at least τ sample points and build a sparse occupancy tree trimming all the nodes with less than τ sample points. Unfortunately, this process can create a sequence of imbedded elements $S_0 \supset S_1 \supset ... \supset S_k$ such that each of the sets $S_j \setminus S_{j+1}$ has less than τ sample points but the cumulative set $S_0 \setminus S_k$ could contain much more than τ sample points. In such a case, we can create a lacunary sequence of indices $i_0 = 0 < i_1 < i_2 < ...$ such that each set $S_{i_j} \setminus S_{i_{j+1}}$ has between, say, τ and 3τ sample points. We then declare these sets elements of the partition and insert them as nodes in the sparse occupancy tree to create an *augmented* sparse occupancy tree with m terminal nodes and the property that the subsets of X, not covered by the elements corresponding to these nodes, have combined measure less than mt. We consider these augmented sparse occupancy trees as the building blocks of the family $\mathcal{S}_m$.

The part (i) of the Theorem holds for the approximation classes $\mathcal{A}^s$ corresponding to the new sequence of families $(\mathcal{S}_m)_{m\geq 1}$. These new classes are much richer than the Besov spaces featured in (ii).

References

[1] P. Binev, A. Cohen, W. Dahmen, R. DeVore, *Classification algorithms using adaptive partitioning*, Ann. Statistics **42** (2014), 2141-2163.

[2] P. Binev, W. Dahmen, and P. Lamby, *Fast high-dimensional approximation with sparse occupancy trees*, J. Comput. Appl. Math. **235** (2011), 2063–2076.

[3] P. Massart and E. Nedelec, *Risk bounds for statistical learning*, Ann. Statistics, **34** (2006), 2326–2366.

[4] A. B. Tsybakov, *Optimal aggregation of classifiers in statistical learning*, Ann. Statistics **32** (2004), 135–166.

Sparse approximation by Prony's method and AAK theory

Gerlind Plonka

(joint work with Vlada Pototskaia)

In signal processing and system theory, we consider the problem of sparse approximation of structured signals. Let us assume that a discrete signal $f := (f_k)_{k=0}^{\infty}$ can be represented by a linear combination of N exponentials,

$$f_k := f(k) = \sum_{j=1}^{N} a_j \, z_j^k \,, \tag{1}$$

where $a_j \in \mathbb{C} \setminus \{0\}$ and $z_j \in \mathbb{D} := \{z \in \mathbb{C} : 0 < |z| < 1\}$. If a suitable number of signal values $f(\ell)$, $\ell = 0, 1, \ldots, M$ with $M \geq 2N - 1$ is given, then the parameters a_j and z_j can be uniquely determined by applying Prony's method, see e.g. [6].

Our goal is now to find a new signal $\tilde{f} := (\tilde{f}_k)_{k=0}^{\infty}$ of the form

$$\tilde{f}_k := \tilde{f}(k) = \sum_{j=1}^{n} \tilde{a}_j \, \tilde{z}_j^k \tag{2}$$

with $\tilde{a}_j \in \mathbb{C} \setminus \{0\}$ and $\tilde{z}_j \in \mathbb{D}$ such that $n < N$ and $\|f - \tilde{f}\|_{\ell^2} \leq \epsilon$.

Problems of this type have been considered already in [3] and [2]. In these papers, an approach using the theory of Adamjan, Arov and Krein [1] has been employed. Furthermore the above approximation problem is strongly related to the problem of structured low rank approximation for Hankel matrices, see e.g. [4]. However, it has been still not completely understood, how to construct the new sequence $\tilde{f}$ in an optimal way.

To solve the above problem, we also employ the AAK theory and consider for the signal of the form (1) the infinite Hankel matrix

$$\mathbf{\Gamma}_f := \begin{pmatrix} f_0 & f_1 & f_2 & \dots \\ f_1 & f_2 & f_3 & \dots \\ f_2 & f_3 & f_4 & \dots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix} = (f_{k+j})_{k,j=0}^{\infty}.$$

Then, it can be simply shown that $\mathbf{\Gamma}_f$ possesses rank N, and we can order the singular values $\sigma_0 \geq \sigma_1 \geq \ldots \geq \sigma_{N-1} > \sigma_N = \ldots = \sigma_\infty = 0$. In particular, $\mathbf{\Gamma}_f$ defines a compact operator on $\ell^2(\mathbb{N}_0)$. For the considered case, a theorem of Adamjan, Arov and Krein [1] states the following.

Theorem 1 (see [1]). *Let f be given as in (1). Further, let $(\sigma_n, u^{(n)})$ with $u^{(n)} = (u_k^{(n)})_{k=0}^{\infty} \in \ell^2(\mathbb{N}_0)$ be a fixed singular pair of $\mathbf{\Gamma}_f$ with $\sigma_n \neq \sigma_k$ for $n \neq k$ and $\sigma_n \neq 0$. Then the series*

$$P_{u^{(n)}}(z) := \sum_{k=0}^{\infty} u_k^{(n)} \, z^k$$

has exactly n zeros $\tilde{z}_1, \ldots, \tilde{z}_n$ in $\mathbb{D}$, repeated according to their multiplicity. Moreover, if $\tilde{z}_1, \ldots, \tilde{z}_n$ are pairwise different, then there exist coefficients $\tilde{a}_1, \ldots, \tilde{a}_n \in \mathbb{C}$ such that for

$$\tilde{f} = (\tilde{f}_j)_{j=0}^{\infty} = \left(\sum_{k=1}^{n} \tilde{a}_k \tilde{z}_k^j \right)_{j=0}^{\infty}$$

we have

$$\|\mathbf{\Gamma}_f - \mathbf{\Gamma}_{\tilde{f}}\| = \sigma_n.$$

The theory behind the above theorem is presented in details in [5]. Note that due to the required structure of $\tilde{f}$ the Hankel matrix $\mathbf{\Gamma}_{\tilde{f}}$ has rank n. Therefore the theorem presents an approach for low rank Hankel approximation, standing in contrast with the approximation by usual singular value decomposition, which doesn't preserve the Hankel structure.

We want to apply this theorem to our sparse approximation problem and will answer the following questions. How ist the operator norm of the Hankel matrix $\mathbf{\Gamma}_f$ related to $\|f\|_{\ell^2}$? How to compute the singular pairs $(\sigma_n, u^{(n)})$ for $n = 0, \ldots, N-1$ numerically? How to find all zeros of the expansion $P_{u^{(n)}}(z)$ lying inside $\mathbb{D}$? How to obtain the optimal coefficients $\tilde{a}_k$?

Using the sequence $e_1 := (1, 0, 0, \ldots)^T \in \ell^2(\mathbb{N}_0)$, it follows that

$$\|f\|_{\ell^2} = \left(\sum_{j=0}^{\infty} |f_j|^2 \right)^{1/2} = \|\mathbf{\Gamma}_f e_1\|_{\ell^2} \leq \sup_{\|u\|_{\ell^2}=1} \|\mathbf{\Gamma}_f u\|_{\ell^2} = \|\mathbf{\Gamma}_f\|.$$

Therefore we have for two sequences f, $\tilde{f} \in \ell^2(\mathbb{N}_0)$ that $\|f - \tilde{f}\|_{\ell^2} \leq \|\mathbf{\Gamma}_f - \mathbf{\Gamma}_{\tilde{f}}\|$.

In order to compute the singular pairs of $\mathbf{\Gamma}_f$ we show the following theorem on the structure of singular vectors resp. con-eigenvectors of $\mathbf{\Gamma}_f$.

Theorem 2. *Let f be of the form (1). Then the con-eigenvectors $u^{(l)} = (u_k^{(l)})_{k=0}^{\infty}$, $l = 0, \ldots, N-1$, corresponding to the nonzero con-eigenvalues $\sigma_0 \geq \ldots \geq \sigma_{N-1} > 0$ of $\mathbf{\Gamma}_f$ are of the form*

$$u_k^{(l)} = \frac{1}{\sigma_l} \sum_{j=1}^{N} a_j P_{\overline{u}^{(l)}}(z_j) z_j^k, \qquad k \in \mathbb{N}_0,$$

where the vectors $(P_{\overline{u}^{(l)}}(z_j))_{j=1}^N = (\overline{P_{u^{(l)}}(\overline{z}_j)})_{j=1}^N$, $l = 0, \ldots, N-1$, are determined by the con-eigenvectors of the finite eigenvalue problem

$$\sigma_l (P_{u^{(l)}}(\overline{z}_j))_{j=1}^N = \mathbf{A}_N \mathbf{Z}_N (\overline{P_{u^{(l)}}(\overline{z}_j)})_{j=1}^N$$

with

$$\mathbf{A}_N := \begin{pmatrix} a_1 & & & 0 \\ & a_2 & & \\ & & \ddots & \\ 0 & & & a_N \end{pmatrix}, \quad \mathbf{Z}_N := \begin{pmatrix} \frac{1}{1-|z_1|^2} & \frac{1}{1-z_1\overline{z}_2} & \cdots & \frac{1}{1-z_1\overline{z}_N} \\ \frac{1}{1-\overline{z}_1 z_2} & \frac{1}{1-|z_2|^2} & \cdots & \frac{1}{1-z_2\overline{z}_N} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{1-\overline{z}_1 z_N} & \frac{1}{1-\overline{z}_2 z_N} & \cdots & \frac{1}{1-|z_N|^2} \end{pmatrix}.$$

Proof. Since $\mathbf{\Gamma}_f$ is symmetric, a singular pair (σ, u) of $\mathbf{\Gamma}_f$ with $u = (u_k)_{k=0}^{\infty}$ is also a con-eigenpair satisfying $\mathbf{\Gamma}_f \overline{u} = \sigma\, u$. Denoting $P_{\overline{u}}(z) := \sum_{k=0}^{\infty} \overline{u}_k z^k$ it follows by (1) that

$$\sigma u_k = (\mathbf{\Gamma}_f \overline{u})_k = \sum_{r=0}^{\infty} f_{k+r} \overline{u}_r = \sum_{r=0}^{\infty} \sum_{j=1}^{N} a_j\, z_j^{k+r}\, \overline{u}_r = \sum_{j=1}^{N} a_j\, P_{\overline{u}}(z_j)\, z_j^k. \tag{3}$$

The assertion of the theorem is now a consequence of (3) and

$$\sigma_l P_{u^{(l)}}(z) = \sigma_l \sum_{r=0}^{\infty} u_r^{(l)} z^r = \sum_{r=0}^{\infty} \sum_{j=1}^{N} a_j P_{\overline{u}^{(l)}}(z_j) z_j^r z^r = \sum_{j=1}^{N} \frac{a_j P_{\overline{u}^{(l)}}(z_j)}{1 - z_j z}$$

for $z \in \mathbb{D}$ by inserting $z = \bar{z}_k$, $k = 1, \ldots, N$. $\square$

From the last equality we observe that $P_{u^{(n)}}(z)$ is a rational function with a numerator being a polynomial of degree at most $N-1$, which enables to compute the zeros of $P_{u^{(n)}}$. Thus the complete algorithm reads as follows.

Algorithm for sparse approximation of exponential sums.

Input: samples f_k, $k = 0, \dots, M$ for sufficiently large $M \geq 2N-1$.
target approximation error ϵ

(1) Find the parameters $z_j \in \mathbb{D}$ and a_j, $j = 1, \dots, N$ of the exponential representation of f in (1) using a Prony-like method.
(2) Solve the con-eigenproblem for the matrix $\mathbf{A}_N \mathbf{Z}_N$ and determine the largest singular value σ_n with $\sigma_n < \epsilon$.
(3) Compute the n zeros $\tilde{z}_j \in \mathbb{D}$ of the con-eigenpolynomial $P_{u^{(n)}}(z)$ of $\mathbf{\Gamma}_f$ using its rational representation.
(4) Compute the coefficients $\tilde{a}_j$ by solving the minimization problem

$$\min_{\tilde{a}_1,\dots,\tilde{a}_n} \|f - \tilde{f}\|_{\ell^2}^2 = \min_{\tilde{a}_1,\dots,\tilde{a}_n} \sum_{k=0}^{\infty} |f_j - \sum_{j=1}^{n} \tilde{a}_j \tilde{z}_j^k|^2.$$

Output: sequence $\tilde{f}$ of the form (2) such that $\|f - \tilde{f}\|_{\ell^2} \leq \sigma_n < \epsilon$.

References

[1] V.M. Adamjan, D.Z. Arov, and M.G. Krein, *Analytic properties of the Schmidt pairs of a Hankel operator and the generalized Schur-Takagi problem*, Mat. Sb. **8**6 (1971), 34–75 (in Russian).
[2] F. Andersson, M. Carlsson, and M.V. de Hoop, *Sparse approximation of functions using sums of exponentials and AAK theory*, J. Approx. Theory **163** (2011), 213–248.
[3] G. Beylkin and L. Monzón, *On approximation of functions by exponential sums*, Appl. Comput. Harmon. Anal. **19** (2005), 17–48.
[4] I. Markovsky, Low Rank Approximation: Algorithms, Implementation, Applications, Springer, London, 2012.
[5] N.K. Nikolski, Operators, Functions, and Systems: An Easy Reading (Mathematical Surveys and Monographs), Vol. **92**, AMS, 2002.
[6] G. Plonka and M. Tasche, *Prony methods for recovery of structured functions*, GAMM-Mitt. **37**(2) (2014), 239–258.

Multiscale radial basis functions: Recent results

Holger Wendland

Radial basis functions are a popular meshfree method. They are used in various areas comprising, for example, scattered data approximation, computer graphics, machine learning, engineering and the geosciences.

Multiscale radial basis functions differ from classical radial basis functions since they do not only use shifts by scattered centres but also different scales. To be more precise, assume that $\Phi : \mathbb{R}^d \to \mathbb{R}$ is a radial basis function with $\Phi(x) = \phi(\|x\|_2)$, $x \in \mathbb{R}^d$, where $\phi : \mathbb{R} \to \mathbb{R}$ is an even function with compact support $[-1, 1]$. Suppose further that we are given a sequence of data sets $X_1, X_2, \dots, X_n \subseteq \Omega \subseteq \mathbb{R}^d$

with decreasing mesh norms $h_j := \sup_{x\in\Omega} \min_{x_j\in X_j} \|x - x_j\|_2$ and a sequence of decreasing support radii $\delta_1 \geq \delta_2 \geq \ldots \geq \delta_n$. Then, we can define local kernels

$$\Phi_j(x, y) = \delta_j^{-d}\Phi((x - y)/\delta_j).$$

and local approximation spaces

$$W_j = \text{span}\{\Phi_j(\cdot, x) : x \in X_j\}.$$

to define global or multiscale approximation spaces

$$V_n := W_1 + \ldots + W_n.$$

While multiscale RBFs have been used for quite some time for pure function reconstruction, but also in the context of solving partial differential equations by collocation, no proofs have been given until very recently.

In my talk, I have addressed recent results on multiscale RBF approximation orders by giving error estimates for a multilevel interpolant. The most recent result states the following.

Assume that Φ is a reproducing kernel of the Sobolev space $H^\sigma(\mathbb{R}^d)$ with $\sigma > d/2$. Assume further that the fill distances and support radii satisfy $h_{j+1} = \mu h_j$ and $\delta_j = \nu h_j$ with $\mu \in (0, 1)$ and $1/h_1 \geq \nu \geq \gamma/\mu$ with $\gamma > 0$ fixed. Then, for every $\epsilon > 0$ there are constants $C > 0$ and $\mu_0 = \mu_0(\epsilon)$ such that

$$\inf_{f_n\in V_n} \|f - f_n\|_{L_2(\Omega)} \leq Ch_n^{\sigma-\epsilon}\|f\|_{H^\sigma(\Omega)} \text{ for all } f \in H^\sigma(\Omega),$$

provided $\mu \leq \mu_0$.

I have discussed a typical residual correction algorithm for computing the multiscale RBF approximation f_n and variations including data compression and adaptivity by showing both theoretical results and examples. Next, I have discussed matrix-valued kernels, which lead to divergence-free approximation spaces and their multiscale extensions. Again, I have discussed convergence orders and gave examples. Finally, I have addressed the topic of using kernel based methods to solve semi-linear parabolic problems on closed, compact, smooth manifolds.

This talk is based on joint work with Quoc Thong Le Gia and Ian Sloan (University of New South Wales, Australia), with Patricio Farrell (WIAS, Germany) and Kathryn Gillow (Oxford University, UK) and on the literature quoted below.

References

[1] P. Farrell, K. Gillow, H. Wendland, Multilevel Interpolation of Divergence-Free Vector Fields, Preprint Berlin/Bayreuth/Oxford, 2015, to appear in: IMA Journal of Numerical Analysis.

[2] Q. T. Le Gia, H. Wendland, Data compression on the sphere using multiscale radial basis functions, Advances in Computational Mathematics 40 (2014), 923–943.

[3] Q. T. Le Gia, I. H. Sloan, H. Wendland, Multiscale analysis in Sobolev spaces on the sphere, SIAM Journal on Numerical Analysis 48 (2010), 2065-2090.

[4] H. Wendland, Multiscale analysis in Sobolev spaces on bounded domains, Numerische Mathematik 116 (2010), 493–517.

[5] H. Wendland, A high-order approximation method for semilinear parabolic equations on spheres, Mathematics of Computation 82 (2013), 227 - 245.

Interpolation with multiquadrics without added constant

Martin Buhmann
(joint work with O. Davydov)

Radial basis function interpolation (and quasi-interpolation) is a useful tool to approximate (at a minimum) continuous real-valued functions f on d-dimensional real space by shifts of translates of a single, radially symmetric function $\phi(\|\cdot\|)$, the norm being usually Euclidean. There are generalisations too to vector- or even matrix-valued functions f.

Apart from the unique existence of such interpolants for all f and distinct interpolations points (so-called "centres"), many different radial functions ϕ and largely independent of the spatial dimension d – which is in itself a highly useful feature – their attractive approximation properties and accuracies even for large d render this approach flexible, useful and suitable for several applications. But while it was noted by Rolland Hardy and proved in a famous paper by Charles A. Micchelli that radial basis function interpolants

$$s(x) = \sum_j \lambda_j \phi(\|x - x_j\|)$$

exist uniquely specifically for arbitrary real parameters c and the multiquadric radial function $\phi(r) = \sqrt{r^2 + c^2}$ in question, as soon as the (at least two) aforementioned centres are pairwise distinct, the achieved *error bounds* for $f(x) - s(x)$ for this interpolation problem *always* demanded an added real constant, call it c, to s.

To make up for this extra degree of freedom, the coefficients λ_j were required to sum to zero. By using Pontryagin native spaces, we obtain attractive error bounds that no longer require this additional real constant expression c and the extra condition on the sum of coefficients; they therefore apply to the original formulation of the interpolants. Some further remarks on quasi-interpolation (joint work with Feng Dai) are added as well.

References

[1] D. Alpay, A. Dijksma, J. Rovnyak, and H.S.V. de Snoo, Reproducing kernel Pontryagin spaces, in "Holomorphic Spaces", Sheldon Axler, John McCarthy, and Donald Sarason (eds.), MSRI Publications Volume 33, Cambridge University Press, 1998, pp. 425–444.

[2] R. Beatson, O. Davydov and J. Levesley, Error bounds for anisotropic RBF interpolation, J. Approx. Theory 162 (2010), 512-527.

[3] G. Berschneider, W. zu Castell, and J. S. Schrödl, Function spaces for conditionally positive definite operator valued kernels, Math. Comp. 81 (2012), 1551–1569.

[4] A. L. Brown, Uniform approximation by radial basis functions, (Appendix B to 'The theory of radial basis functions approximation in 1990' by M.J.D. Powell), in Advances in Numerical Analysis. Vol. II. Wavelets, Subdivision, and Radial Functions, W. A. Light (ed.), Oxford University Press, Oxford, pp. 203–206.

[5] M.D. Buhmann, Radial Basis Functions: Theory and Implementations, Cambridge University Press, 2003.

[6] M.D. Buhmann and F. Dai, Pointwise approximation with quasi-interpolation by radial basis functions, Journal of Approximation Theory 192 (2015), 156–192.

[7] M.D. Buhmann and N. Dyn, Spectral convergence of multiquadric interpolation, Proceedings of the Edinburgh Mathematical Society 36 (1993), 319–333.
[8] O. Davydov, Error bound for radial basis interpolation in terms of a growth function, in "Curve and Surface Fitting: Avignon 2006," (A. Cohen, J.-L. Merrien and L. L. Schumaker, Eds.), Nashboro Press, Brentwood, 2007, pp. 121–130.
[9] O. Davydov and R. Schaback, Error bounds for kernel-based numerical differentiation, Numer. Math. 132 (2016), 243-269.
[10] E. Larsson and B. Fornberg, Theoretical and computational aspects of multivariate interpolation with increasingly flat radial basis functions. Comput. Math. Appl. 49 (2005), 103–130.
[11] C.A. Micchelli, Interpolation of scattered data: distance matrices and conditionally positive definite functions, Constr. Approx. 1 (1986), 11–22.
[12] M.J.D. Powell, Radial basis functions for multivariable interpolation: a review, in Algorithms for Approximation, J.C. Mason and M.G. Cox (eds.), Oxford University Press, Oxford 1987, 143–167.
[13] I.J. Schoenberg, Metric spaces and completely montone functions, Annals Math. 39, 811–841.
[14] Z. Wu and R. Schaback, Local error estimates for radial basis function interpolation of scattered data, IMA J. Numerical Analysis 13 (1993), 13–27.

Causal and statistical learning

Bernhard Schölkopf
(joint work with Dominik Janzing and David Lopez-Paz)

In standard machine learning, the basic object is a joint distribution $P(X)$ generating the observable data. Here, X is a random vector, and we are usually given a dataset $x_1, \ldots, x_n$ sampled i.i.d. from P. We are often interested in estimating properties of conditionals of some components of X given others, e.g., a classifier (which may be obtained by thresholding a conditional at 0.5). This is a nontrivial inverse problem, giving rise to statistical learning theory.

In causal learning in the sense considered here, the basic object is a structural causal model (SCM; also called structure equation model or functional causal model) [2]. In an (acyclic) SCM, the components $X_1, \ldots, X_d$ of X are identified with vertices of a direct acyclic graph whose arrows represent direct causal influences, and there is a noise variable N_i for each vertex, along with a function f_i which computes X_i from its graph parents $\mathrm{Pa}(X_i)$ and N_i, i.e.,

$$X_i = f_i(\mathrm{Pa}(X_i), N_i). \tag{1}$$

The noises N_i are assumed to be jointly independent. The graph connectivity along with the functions then create nontrivial dependences between the observables; moreover, they describe how the system behaves under interventions: by replacing functions (e.g., with constant functions), we can compute the effect of setting some variables to specific values.

The distributions of the noises imply a unique joint distribution of the observables, but since an SCM contains additional information (e.g., about the effect of interventions), the other direction (inferring an SCM from the joint distribution of the observables) is not unique.

In causal structure inference, we seek to infer properties of graph and functions from data. It turns out that subject to certain assumptions, conditional independences among the X_i contain some information about the graph [7]. We have recently shown that assuming a certain type of independence between mechanisms lets us handle some cases that were previously unsolvable [1]; it also has nontrivial implications for machine learning tasks such as semi-supervised learning, covariate shift adaptation and transfer learning [6]. Alternatively, assumptions on the form of the functions, such as additive dependence on the noise, allow us to solve some such cases as well [3].

As mentioned, the SCM description contains more detail; it also allows us to reason about relationships between distributions in more structured ways. If we view an SCM as the underlying generative model for our data, rather than just a joint distribution, then this can have implications for machine learning tasks [5]. We briefly mention some open problems and ideas.

1. Learning multiple tasks in multiple environments. It may be the case that some components of an SCM remain invariant between different learning settings, while others change [6, 4]. This means that even though the settings or environments differ in the joint distributions, some of the components of the distributions' causal factorizations are stable. It is desirable to develop means to identify those components from data [4].

More generally, the different settings may also differ in which predictors are available, and which target variables we are trying to predict. If the settings are deemed related, it would still be desirable to view them as instantiations of different learning problems associated with the same underlying SCM. This is connected to the problem below.

Finally, it may be the case that there is no clear correspondence between predictors in different settings. In this case, one may still be able to train a system (e.g., a neural network) whose components compete for data, get trained on those data where they work best, and thus gradually specialize on different (sub-)tasks, while certain components are shared across tasks.

2. Learning using "privileged information." Whenever we observe data, we think of this as observing a subset of the vertices of an SCM. E.g., we see a histological image x, and we are provided the information y whether there is cancer or not. There are many variables that in reality lie "in between" that we do not see. Sometimes, we get additional features of the image x^* (this is Vapnik's "second space" [8]), possibly computed by a "generator" (or teacher) $p(x^*|x)$, that will make it easier to predict x. "Easier" here could mean that the finite sample error is lower, since by using x^* we get away with a smaller/simpler or more natural function class. Vapnik provides the following example of a generator $p(x^*|x)$: a pathologist who sees the histological image might add the value "aggressive proliferation" for x^*. This kind of metaphorical reasoning is potentially powerful since by this translation we tap into a rich world of shared culture and meaning, assisting us in choosing function classes and generalizing. In the SCM view, we may have seen a lot of data for the mechanism $p(x^*|x)$, but only little data for

$p(y|x)$, and little data for $p(y|x^*)$. In this setting, $p(y|x^*)$ could be easier to learn than $p(y|x)$ if we get away with a simpler function class.

More generally, rather than a teacher $p(x^*|x)$, we might have functions that have been trained on another task, trained on prior data (which is not available for the present task) and evaluated on the current data. For instance, this could be another function in an SCM. Note that in such a setting (where we have different training sets depending on which covariates and which learning task we consider), it may be the case that not only the Markov blanket of the targets should be used for prediction.

3. A compression view of SCMs. In general, there are multiple ways of factorizing a joint distribution into products of conditionals. Different SCMs, be it for all variables or only for subsets, possibly learnt for different (but related) tasks, may or may not share some of those conditionals.

The causal factorization is the one that contains a conditional corresponding to each structural assignment (1). We hypothesize that the causal factorization should permit the most compact overall representation of a collection of SCMs (or of a collection of related training sets), since it will have the largest set of shared conditionals. This may lead to ways of learning causal graphs from multiple datasets and tasks.

Interestingly, this also points to the usefulness of multiple tasks, and more generally:

4. The use of data that is not identically distributed. While traditionally, data that is not i.i.d. has been considered a nuisance, the above indicates that it can be helpful provided the tasks are related (e.g., in a way such that certain conditionals agree). This suggests formalizing such learning tasks. Suppose that we are given an overall set of data points $(x_1, \ldots, x_n)$ (where each x_i could for instance be a pair of inputs and a target); in addition, suppose we have a similarity measure $k(x_i, x_j)$ taking the value 1 if x_i and x_j come from the same distribution, and (close to) 0 if the distributions are very different. How would we best exploit this information to learn an overall set of SCMs?

References

[1] D. Janzing, J. M. Mooij, K. Zhang, J. Lemeire, J. Zscheischler, P. Daniušis, B. Steudel, and B. Schölkopf. Information-geometric approach to inferring causal directions. *Artificial Intelligence*, 182–183:1–31, 2012.

[2] J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, NY, 2nd edition, 2009.

[3] J. Peters, J. M. Mooij, D. Janzing, and B. Schölkopf. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15:2009–2053, 2014.

[4] R. Rojas-Carulla, B. Schölkopf, R. Turner, and J. Peters. Causal transfer in machine learning. *ArXiv e-prints (1507.05333v2)*, 2016.

[5] B. Schölkopf, D. Hogg, D. Wang, D. Foreman-Mackey, D. Janzing, C.-J. Simon-Gabriel, and J. Peters. Modeling confounding by half-sibling regression. *PNAS*, 113(27):7391–7398, 2016.

[6] B. Schölkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. M. Mooij. On causal and anticausal learning. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*, pages 1255–1262, 2012.

[7] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT Press, Cambridge, MA, 2nd edition, 2000.
[8] V. Vapnik and R. Izmailov. Learning using privileged information: Similarity control and knowledge transfer. *Journal of Machine Learning Research*, 16:2023–2049, 2015.

Learning with primal and dual model representations: New extensions

Johan A.K. Suykens

Support vector machines and kernel-based models have been successful for a wide range of problems in supervised and unsupervised learning, and beyond. However, recent developments in deep learning, networks, sparsity and big data are posing new challenges towards a unified understanding, generically applicable methodologies, scalability and new mathematical foundations.

Learning with primal and dual model representations may offer a unifying picture at this point. Primal representations are expressed in terms of the feature map, while dual representations in the kernel function. As previously shown, such characterizations are relevant with respect to sparsity, robustness, out-of-sample extensions, model selection and large scale problems.

In this context, we highlight the following recent extensions:

- Multilevel hierarchical kernel spectral clustering for large scale networks [1]: In kernel spectral clustering a model-based approach is taken to spectral clustering with primal and dual model representations within an optimization problem formulation. One can work with representative subgraphs in a multilevel hierarchical fashion, completing the network through kernel-based out-of-sample extensions. This approach reveals good quality clusters at many scales on real-life networks in comparison with other state-of-the-art methods. For directed networks deformed Laplacians are proposed in [2].

- Generalized support vector regression with tensor-kernel representations [3]: Support vector regression is studied in Banach function spaces using Fenchel-Rockafellar duality. As a result tensor-kernel representations are obtained. This setting of generalized support vector regression admits a larger family of regularization schemes, approaching ℓ^1 regularization, instead of the ℓ^2 regularization that is commonly used in support vector machines.

- New variational principle and nonlinear extensions to singular value decomposition [4]: Matrix singular value decomposition is interpreted within a kernel-based setting with primal and dual model representations. The row and column vectors serve as data sources for which compatible feature maps are considered. The dual problem in the Lagrange multipliers is linked to Lanczos' decomposition theorem. New nonlinear extensions are obtained with general kernels. In the special case of a square symmetric matrix it reduces to a kernel principal component analysis with a Mercer kernel.

- Deep learning using restricted kernel machines and conjugate feature duality [5]: A principle of so-called conjugate feature duality is proposed, which is based on a quadratic form property, proven by the Schur complement. This enables to obtain an interpretation of visible and hidden units for a class of kernel machines (including least squares support vector machines for classification and regression, kernel PCA, matrix SVD, and Parzen-type models), with a dual representation expressed in terms of the hidden features. In this way a connection is made with restricted Boltzmann machines (restricted means here that there are no hidden-to-hidden connections). Deep architectures are obtained by coupling the restricted kernel machines over different levels.

Acknowledgements. Research support by the European Research Council (FP7/2007-2013) / ERC AdG A-DATADRIVE-B (290923); Research Council KUL: GOA/10/09 MaNet, CoE PFV/10/002 (OPTEC), BIL12/11T; PhD/Postdoc grants; Flemish Government: FWO: PhD/ Postdoc grants, projects: G.0377.12 (Structured systems), G.088114N (Tensor based data similarity); IWT: PhD/Postdoc grants, projects: SBO POM (100031); iMinds Medical Information Technologies SBO 2014; Belgian Federal Science Policy Office: IUAP P7/19 (DYSCO).
Publications and preprints available at http://www.esat.kuleuven.be/stadius/ADB/.

References

[1] Mall R., Langone R., Suykens J.A.K., *Multilevel Hierarchical Kernel Spectral Clustering for Real-Life Large Scale Complex Networks*, PLOS ONE, e99966, vol. 9, no. 6, Jun. 2014, pp.1–18.
[2] Fanuel M., Suykens J.A.K., *Deformed Laplacians and spectral ranking in directed networks.* [arXiv:1511.00492]
[3] Salzo S., Suykens J.A.K., *Generalized support vector regression: duality and tensor-kernel representation*, Internal Report 16-62, ESAT-SISTA, KU Leuven (Leuven, Belgium), 2016. [arXiv:1603.05876]
[4] Suykens J.A.K., *SVD revisited: a new variational principle, compatible feature maps and nonlinear extensions*, Applied and Computational Harmonic Analysis, Vol. 40, No. 3, pp. 600–609, May 2016.
[5] Suykens J.A.K., *Deep Restricted Kernel Machines using Conjugate Feature Duality*, Internal Report 16-50, ESAT-SISTA, KU Leuven (Leuven, Belgium), 2016.

Structured high-dimensional estimation

Alexandre Tsybakov
(joint work with Yu Lu, Olga Klopp, Harrison Zhou)

Suppose that we observe a matrix Y satisfying

$$Y = \theta^* + W. \tag{1}$$

Here $\theta^* = (\theta_{ij}) \in \mathbb{R}^{n\times m}$ is the unknown matrix of interest and $W = (W_{ij}) \in \mathbb{R}^{n\times m}$ is the noise matrix. We assume that the signal matrix θ^* is "structured", that is, it can be factorized using sparse factors. Specifically, we assume that

$$\theta^* = (\theta_{ij}) \in \Theta \subset \mathbb{R}^{n\times m}$$

where the class of matrices Θ is defined as

$$\Theta = \{\theta = XBZ^T : X \in \mathcal{A}_{s_n}, B \in \mathbb{R}^{k_n \times k_m}, Z \in \mathcal{A}_{s_m}, \|B\|_\infty \leq B_{\max}\} \tag{2}$$

for some $0 \leq s_n \leq k_n$ and $0 \leq s_m \leq k_m$, where for $r \in \{n, m\}$, we denote by $\mathcal{A}_{s_r}$ either the set containing only the identity $r \times r$ matrix, or

$$\mathcal{A}_{s_r} = \{A \in \mathcal{D}_r^{r \times k_r}, \|A_{i\cdot}\|_0 \leq s_r, \text{ for all } i \in \{1, \ldots, r\}, \|A\|_\infty \leq L\},$$

or

$$\mathcal{A}_{s_r} = \{A \in \mathcal{D}_r^{r \times k_r}, \|A_{i\cdot}\|_0 = s_r, \text{ for all } i \in \{1, \ldots, r\}, \|A\|_\infty \leq L\}.$$

Here, the set $\mathcal{D}_r$ is a subset of $\mathbb{R}$ called an alphabet, the values n, m, k_n, k_m, s_n, s_m are integers, and $B_{\max}, L$ are constants. The notation $A_{i\cdot}$ stands for the ith row of matrix A, $\|b\|_0$ denotes the number of non-zero components of vector b, and $\|A\|_\infty$ is the maximum of components norm of matrix A.

We assume that the noise variables W_{ij} are independent centered sub-Gaussian random variables, i.e., there exists $\sigma > 0$ such that for any $i \in \{1, \ldots, n\}$, $j \in \{1, \ldots, m\}$ we have

$$\mathbb{E} \exp(\lambda W_{ij}) \leq \exp(\lambda^2 \sigma^2 / 2), \quad \forall \lambda > 0.$$

Along with (1), we consider a more general model, which is the matrix completion model. Let $N \leq nm$ be an integer and set $p = \frac{N}{mn}$. We suppose that each entry of Y is observed independently of the other entries with probability p. Let η_{ij} be independent Bernoulli variables with parameter p. Then, the observations are of the form

$$Y_{ij} = \eta_{ij} (\theta_{ij} + W_{ij}). \tag{3}$$

Denote by $Y = (Y_{ij})$ the matrix of observations. The expectation of the number of observed entries is equal to N. This includes as a particular case model (1), which corresponds to $p = 1$ (we observe all entries of Y).

We study the optimal rates of convergence in the Frobenius norm (denoted by $||\cdot||_F$) of estimators of matrix θ^* in a mimimax sense on the class of matrices (2).

This framework is quite general; in specific cases, we obtain several models studied in the literature. Some interesting examples of the corresponding classes Θ are as follows.

- Gaussian mixture:

$$\begin{aligned}\Theta_{GM} = \{&\theta \in \mathbb{R}^{n \times m} : \theta = BZ^T, \text{for some } B \in \mathbb{R}^{n \times k}, \|B\|_\infty \leq B_{\max}, \\ &\text{and } Z \in \{0,1\}^{m \times k} \text{ with } \|Z_{i\cdot}\|_0 = 1\}.\end{aligned}$$

- Sparse Dictionary learning:

$$\begin{aligned}\Theta_{Dict} = \{&\theta \in \mathbb{R}^{n \times m} : \theta = BZ^T, \text{for some } B \in \mathbb{R}^{n \times k}, \|B\|_\infty \leq B_{\max}, \\ &\text{and } Z \in \mathbb{R}^{m \times k} \text{ with } \|Z_{i\cdot}\|_0 \leq s, \|Z\|_\infty \leq L\}.\end{aligned}$$

- Stochastic Block Model (SBM):

$$\begin{aligned}\Theta_{SBM} = \{&\theta \in \mathbb{R}^{n \times n} : \theta = ZBZ^T, \text{for some } B \in [0,1]^{k \times k}, \\ &\text{and } Z \in \{0,1\}^{n \times k} \text{ with } \|Z_{i\cdot}\|_0 = 1\}.\end{aligned}$$

Let $\mathbb{P}_\theta$ denote the distribution of Y satisfying model (3) and let $\mathbb{E}_\theta$ be the expectation with respect to $\mathbb{P}_\theta$. The next theorem gives a minimax lower bound.

Theorem 1. *Let W_{ij} be i.i.d. Gaussian $\mathcal{N}(0,\sigma^2)$ random variables, and let Θ be a class defined in (2). Then there exists a universal constant $c>0$ such that*

$$\inf_{\hat{\theta}} \sup_{\theta\in\Theta} \mathbb{P}_\theta \left\{ ||\hat{\theta}-\theta||_F^2 \geq \frac{c\sigma^2}{p}\left(R_X+R_B+R_Z\right)\right\} \geq 0.7,$$

and

$$\inf_{\hat{\theta}} \sup_{\theta\in\Theta} \mathbb{E}_\theta ||\hat{\theta}-\theta||_F^2 \geq \frac{c\sigma^2}{p}\left(R_X+R_B+R_Z\right),$$

where $R_X = nr_m \wedge ns_n \log\frac{ek_n}{s_n}$, $R_B = r_n r_m$, $R_Z = mr_n \wedge ms_m \log\frac{ek_m}{s_m}$, $r_n = n\wedge k_n$ and $r_m = m\wedge k_m$ provided neither of the sets $\mathcal{A}_{s_n}$ and $\mathcal{A}_{s_m}$ reduces to the identity matrix. If one of them reduces to the identity matrix, the corresponding term R_X or R_Z disappears from the rate. Here, $\inf_{\hat{\theta}}$ denotes the infimum over all estimators.

We conjecture that the lower bound of Theorem 1 gives the minimax rate of convergence for the considered problem. We prove this fact for some specific cases by constructing estimators of θ that achieve this rate. The next theorem shows that the least squares estimator

$$\hat{\theta}^{LS} \in \operatorname*{argmin}_{\theta\in\Theta} ||Y-\theta||_F^2$$

is minimax optimal if $p=1$ and the alphabets D_n and D_m are finite.

Theorem 2. *Let W_{ij} be i.i.d. sub-Gaussian random variables, and let Θ be a class defined in (2) such that D_n and D_m are finite sets. Assume that $k_n \leq n$ and $k_m \leq m$, and $p=1$. Then, there exists a constant $C>0$ depending only on σ, L and $B_{\max}$ such that*

$$\sup_{\theta\in\Theta} \mathbb{E}_\theta ||\hat{\theta}^{LS}-\theta||_F^2 \leq C\left(k_n k_m + ns_n\log\frac{ek_n}{s_n} + ms_m\log\frac{ek_m}{s_m}\right)$$

provided neither of the sets $\mathcal{A}_{s_n}$ and $\mathcal{A}_{s_m}$ reduces to the identity matrix. If one of them reduces to the identity matrix, the corresponding term $ns_n\log\frac{ek_n}{s_n}$ or $ms_m\log\frac{ek_m}{s_m}$ disappears from the rate.

As a consequence of Theorems 1 and 2, we obtain the optimal rates of convergence for the Gaussian mixture and SBM models, for which the sets D_n and D_m have cardinality 2.

Corollary 1. *Let W_{ij} be i.i.d. Gaussian $\mathcal{N}(0,\sigma^2)$ random variables, and $p=1$. Then,*

$$\inf_{\hat{\theta}} \sup_{\theta\in\Theta_{GM}} \mathbb{E}_\theta ||\hat{\theta}-\theta||_F^2 \asymp nk + m\log k,$$

$$\inf_{\hat{\theta}} \sup_{\theta\in\Theta_{SBM}} \mathbb{E}_\theta ||\hat{\theta}-\theta||_F^2 \asymp k^2 + n\log k,$$

provided that $k\leq m$ for Θ_{GM} and $k\leq n$ for Θ_{SBM}.

For the general scheme of matrix completion, we obtain the following upper bound, which departs from the lower bound of Theorem 1 by a logarithmic factor.

Theorem 3. *Let W_{ij} be i.i.d. sub-Gaussian random variables, and let Θ be a class defined in (2). Assume that there exists a constant $0 < c' < 1$ such that $\log(s_n) \leq c' \log(k_n)$ and $\log(s_m) \leq c' \log(k_m)$. Then there exists a constant $C > 0$ depending only on σ, L and $B_{\max}$ such that*

$$\sup_{\theta \in \Theta} \mathbb{E}_\theta ||\hat{\theta}^{LS} - \theta||_F^2 \leq \frac{C \log(m \vee n)}{p} \left(R_X + R_B + R_Z\right),$$

provided neither of the sets $\mathcal{A}_{s_n}$ and $\mathcal{A}_{s_m}$ reduces to the identity matrix. If one of them reduces to the identity matrix, the corresponding term R_X or R_Z disappears from the rate.

Iterative multiplicative filters for data labeling

Gabriele Steidl

(joint work with Ronny Bergmann, Jan Henrik Fitschen and Johannes Persch)

Data labeling is a basic problem which appears in many applications. In particular it can be used for image partitioning and segmentation, which is an important preprocessing step for many state-of-the-art algorithms used for performing high-level computer vision tasks. A huge number of different methods has been developed for this purpose and no single technique works best for all cases.

Recently, Åström, Petra, Schmitzer, and Schnörr [1] suggested an interesting supervised geometric approach to the labeling problem. The objective function to minimize is defined on the manifold of stochastic matrices and a minimizing algorithm via the corresponding Riemannian gradient ascent flow is considered. In the numerical part the authors apply several simplifications, in particular lifting maps which finally lead to a simple iterative procedure. Unlike the continuous Riemannian gradient flow that is shown in [1] to converge to unambiguous labelings, the authors merely showed that the simplified numerical scheme closely approximates this flow, but did not prove its convergence. This proof is implied by our results reported in the present paper.

We propose a simple algorithm that can be seen as an iterative multiplicative filtering of a label assignment matrix which can be used to assign K labels to $n \gg K$ data points. Then the i-th row W_i^T of a label assignment matrix $W \in \mathbb{R}^{n,K}$ contains a vector in the probability simplex whose k-th entry gives the probability that the i-th data point belongs to label $k \in \{1, \ldots, K\}$. We start the iterations with a label assignment matrix containing the averaged distances between the prior data and the observed ones. Here the data and their priors may lie in any metric space, which makes the method highly flexible. Then this label assignment matrix is iterated in a multiplicative way. We prove that under mild conditions the iterates in each row converge to unit vectors in $\mathbb{R}^K$, i.e., to vertices of the probability simplex. This enables a unique label assignment. Clearly, our filters

are no smoothing filters, but in contrary force the rows of the label assignment matrix to move to the vertices of the probability simplex and therefore to minimize their entropy. We show that from another point of view each iteration can be understood as finding weighted barycenters of the previous iterates with respect of the Kullback-Leibler distance on the probability simplex. Our filters can be also of non-local nature, but are geometric means instead of arithmetic ones.

A modification of our algorithm resembles the original idea of Åström, Petra, Schmitzer, and Schnörr [1]. Since the relation is not immediately clear, we provide the corresponding details. Further, we add a convergence result for the modified method.

Numerical results demonstrate the very good performance of our algorithm. In particular we apply the method for the partitioning of manifold-valued images as $SO(3)$ valued EBSD data, see also [2]. For more information we refer to [3]

References

[1] F. Åström, S. Petra, B. Schmitzer, and C. Schnörr. Image labeling by assignment. *ArXiv Preprint 1603.05285*, 2016.

[2] R. Bergmann, R. H. Chan, R. Hielscher, J. Persch, and G. Steidl. Restoration of manifold-valued images by half-quadratic minimization. Inverse Problems in Imaging 10(2), 281–304, 2016.

[3] R. Bergmann, J. H. Fitschen, J. Persch and G. Steidl. Iterative multiplicative filters for data labeling. ArXiv Preprint 1604.08714, 2016.

Learning action potential dynamics for preclinical drug safety testing

Philipp Kügler

The term action potential (AP) refers to the characteristic membrane voltage response of excitable cells such as cardiomyocytes to a superthreshold electric stimulus. Cardiac APs underly the contraction of the myocardium and are regulated by a subtle interplay of various ion channels that control the in- and outflow of ions across the membrane. If this interplay is perturbed by pharmaceutical compounds, the AP gets impaired and arrhythmias such as early afterdepolarizations (EADs) may arise. These pathological voltage oscillations may synchronize at the tissue level and trigger potentially lethal ventricular fibrillation.

While drug cardiotoxicity is of major concern both to the pharmaceutical industry and regulatory agencies, the current preclinical in vitro and animal assays for predicting proarrhythmic risk are recognized deterrents to drug development due to lack of specificity. As a consequence, the US FDA recently triggered the CiPA initiative for a radical overhaul of the drug safety paradigm. Therein, breakthrough is expected from a combination of human stem cell technology with mathematical modelling of cardiac action potentials.

Cardiac action potentials are mathematically described by means of coupled systems of nonlinear ODEs that consider the cellular membrane as an electrical

circuit consisting of a capacitative current in parallel with several transmembrane ionic currents. Therein, the voltage equation

$$C\frac{dV}{dt} = -\sum_{ion} I_{ion} + I_{sti} \tag{1}$$

is complemented by additional ODEs for channel gating variables that describe the voltage dependent activation and deactivation of the ionic currents. This modelling approach dates back to the work of Denis Noble [2] and forms the basis of all modern models for animal, human adult and human pluripotent stem cell derived cardiomyocytes. For studying the impact of pharmaceutical compounds on the action potential, (1) then is extended by models of drug-ion channel interaction [4].

Applying multiple time scale analysis and bifurcation theory to cardiac action potential models (1), early afterdepolarizations have been attributed with different bifurcations of equilibria in the fast AP subdynamics [1], [5]. Likewise, arrhythmic behaviour can be associated with bifurcations of limit cycles in full time scale AP models in which corresponding bifurcation parameters can be interpreted in terms of drug action. Consequently, distances to bifurcations can be utilized as safety margin for preclinical drug cardiotoxicity testing and offer a promising supplement to non-mechanistic logistic regression based risk estimation.

In that regard, one challenge is that experimentally validated drug-AP models are not readily available at the early stages of drug development. Both the heterogeneity of cardiomyocytes even of the same type and the lack of a full pharmacological characterization of a new compound to be tested hamper the use of only slightly adapted drug-AP models from the shelf. On the other hand, the limited amount of experimental and computational efforts that can be devoted at the preclinical stage to a single compound the repetition of otherwise in principle straightforward model building and validation steps.

Against this background we suggest the use of learning algorithms [6] for deriving parsimonious AP model equations directly from experimental data, that is to identify the vector valued function f in

$$\frac{dz}{dt} = f(z).$$

Collecting time series data of z and $\dot{z}$ as

$$\mathbf{Z} = \begin{pmatrix} z_1(t_1) & z_2(t_1) & \dots & z_n(t_1) \\ z_1(t_2) & z_2(t_2) & \dots & z_n(t_2) \\ \vdots & \vdots & \ddots & \vdots \\ z_1(t_m) & z_2(t_m) & \dots & z_n(t_m) \end{pmatrix}, \dot{\mathbf{Z}} = \begin{pmatrix} \dot{z}_1(t_1) & \dot{z}_2(t_1) & \dots & \dot{z}_n(t_1) \\ \dot{z}_1(t_2) & \dot{z}_2(t_2) & \dots & \dot{z}_n(t_2) \\ \vdots & \vdots & \ddots & \vdots \\ \dot{z}_1(t_m) & \dot{z}_2(t_m) & \dots & \dot{z}_n(t_m) \end{pmatrix},$$

and constructing a library $\Theta(\mathbf{Z})$ of candidate functions of the columns of $\mathbf{Z}$, e.g.,

$$\Theta(\mathbf{Z}) = \left[\mathbf{1} \,|\, \mathbf{Z} \,|\, \mathbf{Z^{P_2}} \,|\, \cdots \,|\, \sin(\mathbf{Z}) \,|\, \cdots\right]$$

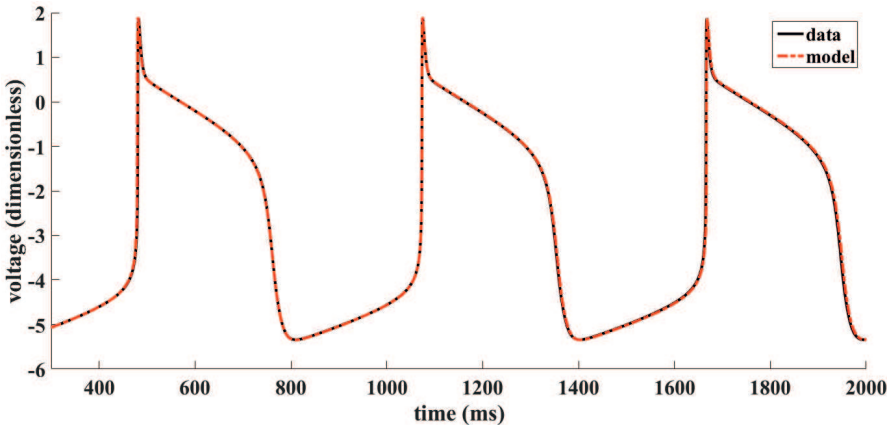

FIGURE 1. Sparse reconstruction of action potential dynamics

with

$$\mathbf{Z^{P_2}} = \begin{pmatrix} z_1^2(t_1) & z_1(t_1)z_2(t_1) & \dots & z_2^2(t_1) & \dots & z_n^2(t_1) \\ z_1^2(t_2) & z_1(t_2)z_2(t_2) & \dots & z_2^2(t_2) & \dots & z_n^2(t_2) \\ \vdots & \vdots & \ddots & \vdots\vdots & \ddots & \vdots \\ z_1^2(t_m) & z_1(t_m)z_2(t_m) & \dots & z_2^2(t_m) & \dots & z_n^2(t_m) \end{pmatrix},$$

the goal is to determine the vectors of coefficients

$$\mathbf{Q} = [\mathbf{q_1}\ \mathbf{q_2}\ \cdots\ \mathbf{q_n}]$$

in

$$\dot{\mathbf{Z}} = \Theta(\mathbf{Z})\mathbf{Q}.$$

Under the assumption that only a few relevant terms define the dynamics (such that f is sparse in a high-dimensional nonlinear functions space), sparse regression can be used for solving the inverse problem. With $\Theta(\mathbf{z^T})$ as vector of symbolic functions, the identified model then is

$$\frac{dz}{dt} = f(z) = \mathbf{Q^T}(\Theta(\mathbf{z^T}))^{\mathbf{T}} \tag{2}$$

In context of preclinical drug cardiotoxicity testing, our suggestion/goal is to derive a parsimonious drug-AP model (2) in an automated manner for a test compound at hand and subsequently determine its distance from the closest bifurcation associated with arrhythmic behaviour.

In a first step, we tested the reconstruction of the action potential dynamics of a polynomial approximation [3] to the original Noble model. The Figure 1 displays the sparse reconstruction of the transmembrane voltage using a candidate library of 120 mulitvariate monomials of degree ≤ 7, where only 25 out of 360 model parameters were determined to be non-zero. Future challenges include the consideration of noisy data, partial state observations and the use of non-polynomial basis functions possibly more appropriate for the description of AP dynamics.

References

[1] D.X. Tran, D. Sato, A. Yochelis, J.N. Weiss, A. Garfinkel, Z. Qu, *Bifurcation and Chaos in a Model of Cardiac Early Afterdepolarizations*, Physical Review Letters **102** (2009), 258103.

[2] D. Noble, *A modification of the Hodgkin-Huxley equations applicable to Purkinje fibre action and pacemaker potentials*, The Journal of Physiology **160** (1962), 317–352.

[3] G. Duckett, D. Barkley, *Modeling the Dynamics of Cardiac Action Potentials*, Physical Review Letters **85** (2000), 884–887.

[4] T. Brennan, M. Fink, B. Rodriguez, *Multiscale modelling of drug-induced effects on cardiac electrophysiological activity*, European Journal of Pharmaceutical Sciences **36** (2009), 62–77.

[5] P. Kügler, *Early afterdepolarizations with growing amplitudes via delayed subcritical Hopf bifurcations and unstable manifolds of saddle foci in cardiac action potential dynamics*, PLoS ONE 11(3): e0151178 (2016).

[6] S.L. Brunton, J.L. Proctor, J.N. Kutz, *Discovering governing equations from data by sparse identification of nonlinear dynamical systems*, PNAS **113** (2016), 3932–3937.

Distributed learning algorithms

Ding-Xuan Zhou

Distributed learning based on a divide-and-conquer approach has the advantages of reducing the memory and computing costs to handle big data. This method has been observed to be very successful in many practical applications. A distributed learning algorithm consists of three steps: partitioning the data into disjoint subsets, applying a learning algorithm implemented in an individual machine or processor to each data subset to produce an individual output, and synthesizing a global output by utilizing some average of the individual outputs. In this talk we discuss error analysis of some distributed learning algorithms including least squares regularization schemes and spectral algorithms.

Let X be a compact metric space (input space), $Y = \mathbb{R}$ (output space) and ρ be a probability measure on $X \times Y$. Take a random sample $D = \{(x_i, y_i)\}_{i=1}^N$ independently drawn from ρ. The regression function f_ρ is defined by $f_\rho(x) = \int_Y y d\rho(y|x)$ where $\rho(\cdot|x)$ is the conditional distribution of ρ at $x \in X$.

We consider learning for regression based on a Mercer kernel $K : X \times X \to \mathbb{R}$ which is a continuous, symmetric and positive semi-definite function generating a reproducing kernel Hilbert space (RKHS) $(\mathcal{H}_K, \|\cdot\|_K)$ by functions $\{K_x = K(\cdot, x) : x \in X\}$. The approximation ability of the RKHS for kernel methods may be measured by the range of L_K^r, the r-th power with $r > 0$ of the integral operator L_K on $L^2_{\rho_X}$ defined by

$$L_K(f)(x) = \int_X K(x, y) f(y) d\rho_X(y), \qquad x \in X,$$

where ρ_X is the marginal distribution of ρ on X. The complexity of the RKHS may be measured by the effective dimension defined as the trace of the operator $(L_K + \lambda I)^{-1} L_K$

$$\mathcal{N}(\lambda) = \mathrm{Tr}\left((L_K + \lambda I)^{-1} L_K\right) = \sum_i \frac{\lambda_i}{\lambda_i + \lambda}, \qquad \lambda > 0,$$

where λ_i is the i-th eigenvalue of L_K.

The first algorithm we discuss [1, 2] is distributed learning with the least squares regularization scheme defined by

$$f_{D,\lambda} = \arg\min_{f\in\mathcal{H}_K} \left\{ \frac{1}{N}\sum_{i=1}^{N}(f(x_i)-y_i)^2 + \lambda\|f\|_K^2 \right\}, \quad \lambda > 0. \tag{1}$$

If we divide the sample $D = \{(x_i, y_i)\}_{i=1}^N$ of input-output pairs into disjoint subsets $\{D_j\}_{j=1}^m$, applying a learning algorithm to the much smaller data subset D_j gives an output f_{D_j}, and the global output might be $\overline{f}_D = \frac{1}{m}\sum_{j=1}^m f_{D_j}$. Here $f_{D_j} = f_{D_j,\lambda}$ and $\overline{f}_D = \overline{f}_{D,\lambda}$.

Theorem 1. Assume $|y| \le M$ almost surely and $f_\rho = L_K^r(g_\rho)$ for some $0 \le r \le \frac{1}{2}$ and $g_\rho \in L^2_{\rho_X}$. If $\mathcal{N}(\lambda) = O(\lambda^{-\frac{1}{2\alpha}})$ for some $\alpha > 0$, $|D_j| = \frac{N}{m}$ for $j = 1, \dots, m$, and $m \le N^{\min\{\frac{12\alpha r+1}{5(4\alpha r+2\alpha+1)}, \frac{4\alpha r}{4\alpha r+2\alpha+1}\}}$, then by taking $\lambda = N^{-\frac{2\alpha}{4\alpha r+1}}$, we have

$$E\left[\left\|\overline{f}_{D,\lambda} - f_\rho\right\|_{L^2_{\rho_X}}\right] = O\left(N^{-\frac{\alpha+2\alpha r}{2\alpha+4\alpha r+1}}\right).$$

If $f_\rho \in \mathcal{H}_K$ and $m \le N^{\frac{1}{4+6\alpha}}$, the choice $\lambda = \left(\frac{m}{N}\right)^{\frac{2\alpha}{2\alpha+1}}$ yields

$$E\left[\left\|\overline{f}_{D,\lambda} - f_{D,\lambda}\right\|_{L^2_{\rho_X}}\right] = O\left(N^{-\frac{\alpha}{2\alpha+1}} m^{-\frac{1}{4\alpha+2}}\right)$$

and

$$E\left[\left\|\overline{f}_{D,\lambda} - f_{D,\lambda}\right\|_K\right] = O\left(\frac{1}{\sqrt{m}}\right).$$

The second algorithm we discuss [3] is distributed learning with spectral algorithms

$$f_{D,\lambda} = g_\lambda(L_{K,D})\frac{1}{N}\sum_{i=1}^{N} y_i K_{x_i}$$

induced by a filter function $g_\lambda : [0, \kappa^2] \to \mathbb{R}$ with a parameter $\lambda > 0$ and the empirical integral operator $L_{K,D}$ associated with the kernel K and the input data $D(\mathbf{x}) = \{x_1, \cdots, x_N\}$ defined on $\mathcal{H}_K$ as

$$L_{K,D}(f) = \frac{1}{N}\sum_{i=1}^{N} f(x_i)K_{x_i} = \frac{1}{N}\sum_{i=1}^{N}\langle f, K_{x_i}\rangle_K K_{x_i}.$$

Here $\kappa = \max_{x\in X}\sqrt{K(x,x)}$ and $g_\lambda(L_{K,D})$ is a linear operator on $\mathcal{H}_K$ defined by spectral calculus: if $\{(\sigma_i^{\mathbf{x}}, \phi_i^{\mathbf{x}})_i\}$ is a set of normalized eigenpairs of $L_{K,D}$, then $g_\lambda(L_{K,D}) = \sum_i g_\lambda(\sigma_i^{\mathbf{x}})\phi_i^{\mathbf{x}} \otimes \phi_i^{\mathbf{x}} = \sum_i g_\lambda(\sigma_i^{\mathbf{x}})\langle\cdot, \phi_i^{\mathbf{x}}\rangle_K \phi_i^{\mathbf{x}}$. Spectral algorithms have been well developed in the literature of inverse problems and learning theory [4, 5]. Examples include Landweber iteration, spectral cut-off, and accelerated Landweber iteration. The least squares regularization is a special spectral algorithm with $g_\lambda(\sigma) = (\sigma + \lambda)^{-1}$. It has a saturation phenomenon in the sense that its approximation ability does not increase when the regularity of the approximated function goes beyond certain level. Spectral algorithms can be used to overcome

the saturation when the filter function $g_\lambda : [0, \kappa^2] \to \mathbb{R}$ with $0 < \lambda \leq \kappa^2$ has a large qualification defined as a positive number ν_g such that there exists a positive constant b independent of λ satisfying

$$\sup_{0<\sigma\leq\kappa^2} |g_\lambda(\sigma)| \leq \frac{b}{\lambda}, \qquad \sup_{0<\sigma\leq\kappa^2} |g_\lambda(\sigma)\sigma| \leq b,$$

and

$$\sup_{0<\sigma\leq\kappa^2} |1 - g_\lambda(\sigma)\sigma|\sigma^\nu \leq \gamma_\nu \lambda^\nu, \qquad \forall\ 0 < \nu \leq \nu_g,$$

where $\gamma_\nu > 0$ is a constant depending only on $\nu \in (0, \nu_g]$.

Theorem 2. Assume $|y| \leq M$ almost surely and the filter function $g_\lambda : [0, \kappa^2] \to \mathbb{R}$ with $0 < \lambda \leq \kappa^2$ has a qualification $\nu_g \geq \frac{1}{2}$. If $f_\rho = L_K^r(g_\rho)$ for some $\frac{1}{2} \leq r \leq \nu_g$ and $g_\rho \in L^2_{\rho_X}$, $\mathcal{N}(\lambda) = O(\lambda^{-\beta})$ for some $\beta > 0$, $|D_j| = \frac{N}{m}$ for $j = 1, \ldots, m$, $\lambda = N^{-\frac{1}{2r+\beta}}$, and

$$m \leq N^{\min\left\{\frac{2}{2r+\beta}, \frac{2r-1}{2r+\beta}\right\}},$$

then

$$E[\|\overline{f}_{D,\lambda} - f_\rho\|^2_{L^2_{\rho_X}}] = \mathcal{O}\left(N^{-\frac{2r}{2r+\beta}}\right).$$

References

[1] Y. C. Zhang, J. Duchi, and M. Wainwright, Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates, Journal of Machine Learning Research **16** (2015), 3299-3340.

[2] S. Lin, X. Guo, and D. X. Zhou, Distributed learning with least square regularization, manuscript, 2015.

[3] Z. C. Guo, S. B. Lin, and D. X. Zhou, Learning theory of distributed spectral algorithms, Inverse Problems, minor revision under review, 2016.

[4] H. W. Engl, M. Hanke, and A. Neubauer, Regularization of Inverse Problems, Vol. 375 of Mathematics and Its Applications, Kluwer Academic Publishers, 1996, Dordrecht.

[5] L. Lo Gerfo, L. Rosasco, F. Odone, E. De Vito, and A. Verri, Spectral algorithms for supervised learning, Neural Computation **20** (2008), 1873-1897.

Learning with hierarchical kernels

Ingo Steinwart

(joint work with Philipp Thomann)

Although kernel methods such as support vector machines are one of the state-of-the-art methods when it comes to fully automated learning, see e.g. the recent independent comparison [2], the recent years have shown that on complex datasets such as image, speech and video data, they clearly fall short compared to deep neural networks.

One possible explanation for this superior behavior is certainly their deep architecture that makes it possible to represent highly complex functions with relatively few parameters. In particular, it is possible to amplify or surpress certain dimensions or features of the input data, or to combine features to new, more abstract

features. Compared to this, standard kernels such as the popular Gaussian kernels simply treat every feature equally. In addition, most users of kernel machines probably stick to the very few standard kernels, often simply because there is in most cases no principled way for finding problem specific kernels. In contrast to this, deep neural networks offer yet another order of freedom to the user by making it easy to choose among many different architectures and other design decisions. This discussion shows that the class of deep neural networks offers potentially much more functions that may fit well to the problem at hand than classical kernel methods do. Therefore, if the training algorithms are able to find these good hypotheses while simultaneously controlling the inherent danger of overfitting (and the user picked a good design), then the recent success of deep networks does not seem to be so surprising after all. In particular, this may be an explanation for the types of data mentioned above, for which an equal and un-preprocessed use of all features may really not be the best idea. Moreover, the recent success of deep networks indicates that these 'ifs' can nowadays much better be controlled than 20 to 30 years ago. This naturally raises the question, whether and how certain aspects of deep neural networks can be translated into the kernel world without sacrificing the benefits of kernel-based learning, namely less 'knobs' an unexperienced user can play with, the danger of getting stuck in poor local minima, or more principled statistical understanding, and last but not least, their success in situations in which no human expert is in the loop.

Of course, the limitations of using simple single kernels have been recognized before. Probably the first attempts in this direction are multiple kernel learning algorithms, see [3], which, in a nutshell, replace a single kernel by a weighted sum of kernels. The advantage of this approch is that finding these weights can again be formulated as a convex objective, while the disadvantage is the limited gain in expressive power unless the used dictionary of kernels is really huge. A more recent approach for increasing the expressive power is to construct complex kernels from simple ones by composing their feature maps in some form. Probably the first result in this direction can be found in [1], in which the authors described the general setup and considered some particular constructions. Moreover, this idea was adopted in [4], where the authors considered sums of kernels in each composition step and established bounds on the Rademacher chaos complexities. Similarly, [7] considered such sums in the decomposition step, but they mostly restrict their considerations to a single decomposition step, for which they establish a generalization bound based on the pseudo-dimension. Furthermore, [6] investigate compositions, in which the initial map is not a kernel feature map, but the map induced by a deep network. All these articles also present some experimental results indicating the benefits of the more expressive kernels. Finally, [5] reports some experiments with a linear support vector machine (SVM) as the top layer of a deep network.

In this talk we also adopt the idea of iteratively composing weighted sums of kernels in each layer. Unlike the papers mentioned above, however, we focus on sums of Gaussian kernels composed with Gaussian kernels. To be more precise,

consider a kernel of the form

$$k_{\gamma,X,H}(x,x') := \exp\bigl(-\gamma^{-2}||\Phi(x)-\Phi(x')||_H^2\bigr)\,, \qquad x,x' \in X\,, \tag{1}$$

where $X \subset \mathbb{R}^d$, H is a Hilbert space, and $\Phi : X \to H$ is some map. Now note that $k(x,x') := \langle \Phi(x), \Phi(x')\rangle$ defines a kernel, and since we have $||\Phi(x)-\Phi(x')||_H^2 = k(x,x) - 2k(x,x') + k(x',x')$, we can also express (1) by the kernel k. This observation in particular applies to kernels k of the form

$$k(x,x') := \sum_{i=1}^{l} w_i^2 k_i\bigl(x_{I_i}, x'_{I_i}\bigr)\,, \qquad x,x' \in X_I. \tag{2}$$

where $I_1,\ldots,I_l \subset \{1,\ldots,d\}$, $x_{I_i} := (x_j)_{j\in I_i}$ denotes a projected vector, k_i are kernels on $X_{I_i} := \{x_{I_i} : x \in X\}$ and $w_1,\ldots,w_l > 0$ are weights. This leads to:

Definition Let k be a kernel of the form (2) and H be its reproducing kernel Hilbert space (RKHS). Then the resulting kernel $k_{\gamma,X_I,H}$ is said to be a hierarchical Gaussian kernel

(1) of depth 1, if all kernels $k_1,\ldots,k_l$ in (2) are linear kernels.
(2) of depth $m > 1$, if all $k_1,\ldots,k_l$ in (2) are hierarchical Gaussian kernels of depth $m-1$.

Besides an illustrative interpretation of the construction, which highlights the similarities to deep architectures, we show the following approximation result:

Theorem *Let $X \subset \mathbb{R}^d$ be compact and $I = \{1,\ldots,d\}$. Then every hierarchical Gaussian kernel $k_{\gamma,X_I,H}$ of some depth is universal, i.e. its RKHS is dense in $C(X)$.*

Based on this result we further show that the corresponding SVMs become universally consistent. We then describe an optimization algorithm for finding the kernel weights, and last but not least we report results from extensive experiments comparing different architectures against each other and against a vanilla SVM with Gaussian kernel. It turns out that even with very moderately sized networks such as depth 2 kernels with four or eight kernels $k_1,\ldots,k_l$ in (2) the standard SVM can consistently be outperformed, sometimes with very significant improvements. In addition, we investigate empirically, whether the learned kernel can also be used for other purposes.

References

[1] Y. Cho and L.K. Saul. Kernel methods for deep learning. In Y. Bengio, D. Schuurmans, J.D. Lafferty, C.K.I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 342–350. 2009.

[2] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim. Do we need hundreds of classifiers to solve real world classification problems? *J. Mach. Learn. Res.*, 15:3133–3181, 2014.

[3] S. Sonnenburg, G. Rätsch, C. Schäfer, and B. Schölkopf. Large scale multiple kernel learning. *J. Mach. Learn. Res.*, 7:1531–1565, 2006.

[4] E.V. Strobl and S. Visweswaran. Deep multiple kernel learning. In *Machine Learning and Applications (ICMLA), 2013 12th International Conference on*, volume 1, pages 414–417, 2013.
[5] Yichuan Tang. Deep learning using support vector machines. In *ICML 2013 Challenges in Representation Learning Workshop*, 2013.
[6] A.G. Wilson, Z. Hu, R. Salakhutdinov, and E.P. Xing. Deep kernel learning. *JMLR W&CP*, 51:370–378, 2016.
[7] J. Zhuang, I.W. Tsang, and S. Hoi. Two-layer multiple kernel learning. *JMLR W&CP*, 15:909–917, 2011.

Learning dynamical systems

Sayan Mukherjee
(joint work with Kevin McGoff, Andrew Nobel, Natesh Pillai)

We consider the asymptotic consistency of maximum likelihood parameter estimation for dynamical systems observed with noise. Under suitable conditions on the dynamical systems and the observations, we show that maximum likelihood parameter estimation is consistent. Our proof involves ideas from both information theory and dynamical systems. Furthermore, we show how some well-studied properties of dynamical systems imply the general statistical properties related to maximum likelihood estimation. Finally, we exhibit classical families of dynamical systems for which maximum likelihood estimation is consistent. Examples include shifts of finite type with Gibbs measures and Axiom A attractors with SRB measures.

We also develop a relative version of the thermodynamic formalism and investigate its connections to Bayesian inference.We consider Bayesian inference procedures in the setting of ergodic observations. By developing a general theory for the asymptotic analysis of such procedures, we will generalize the classical thermodynamic formalism to the relative setting. Then we will we apply our theoretical tools to establish rigorous results on the consistency of common Bayesian inference procedures involving Gibbs measures.

References

[1] K. McGoff, S. Mukherjee, A. Nobel, N. Pillai, *Consistency of maximum likelihood estimation for some dynamical systems*, Annals of Statistics **43**:1 (2015), 1–29.

Robust pairwise learning with kernels

Andreas Christmann
(joint work with Ding-Xuan Zhou)

Regularized empirical risk minimization plays an important role in machine learning theory. A broad class of regularized pairwise learning (RPL) methods based on kernels is investigated. One example is regularized minimization of the error entropy loss which has recently attracted quite some interest from the viewpoint

of consistency and learning rates. We show that such RPL methods have additionally good statistical robustness properties, if the loss function and the kernel are chosen appropriately. The talk is based on [2].

Let $(\mathcal{X}, \mathcal{A})$ be a measurable space, $\mathcal{Y} \subset \mathbb{R}$ be closed, $\mathcal{B}_\mathcal{Y}$ the Borel σ-algebra on $\mathcal{Y}$, and P be a probability measure on $(\mathcal{A} \otimes \mathcal{B}_\mathcal{Y})$. Then a function $L : (\mathcal{X} \times \mathcal{Y})^2 \times \mathbb{R}^2 \to [0, \infty)$ is called a pairwise loss function, if it is measurable. A pairwise loss function L is called separately Lipschitz continuous, if there exists a constant $|L|_1 \in [0, \infty)$ such that, for all $t, \tilde{t}, t', \tilde{t}' \in \mathbb{R}$,

$$\sup_{x,\tilde{x}\in\mathcal{X},y,\tilde{y}\in\mathcal{Y}} \left|L(x, y, \tilde{x}, \tilde{y}, t, \tilde{t}) - L(x, y, \tilde{x}, \tilde{y}, t', \tilde{t}')\right| \le |L|_1 \left(|t - t'| + |\tilde{t} - \tilde{t}'|\right).$$

is satisfied. If L is a pairwise loss function, then $L^\star := L - L(\cdot, \cdot, \cdot, \cdot, 0, 0)$ is called shifted pairwise loss function. To consider shifted loss functions is essential for robustness considerations to reduce moment conditions. If L is a separately Lipschitz pairwise loss functions, no moment conditions on P at all are needed for our results. Hence, even heavy tailed probability measures are allowed, e.g. Cauchy distribution, stable distributions, or mixtures of Gaussian distributions with Cauchy distributions. This is not possible e.g. for the least squares pairwise loss function.

Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a kernel with reproducing kernel Hilbert space H and $\Phi(x) := k(\cdot, x)$, $x \in \mathcal{X}$, its reproducing kernel Hilbert space. The RPL operator S maps any Borel probability measure P on $(\mathcal{X} \times \mathcal{Y}, \mathcal{B}_{\mathcal{X}\times\mathcal{Y}})$ to

$$S(\mathrm{P}) := f_{L^\star,\mathrm{P},\lambda} := \arg\inf_{f\in H} \mathsf{E}_{\mathrm{P}\otimes\mathrm{P}} L^\star(X, Y, \tilde{X}, \tilde{Y}, f(X), f(\tilde{X})) + \lambda\|f\|_H^2$$

and the RPL estimator is defined by

$$S_n((X_1, Y_1), \ldots, (X_n, Y_n)) = S(\mathbb{P}_n),$$

where $\mathbb{P}_n$ denotes the empirical measure.

We first derive results on existence and uniqueness. Our main tool for our robustness results is a new representer theorem, which covers convex and non-convex pairwise loss functions.

If the pairwise loss function is bounded and non-convex, then we derive an upper bound for the maximum bias of the regularized risk in total variation or contamination neighborhoods. This upper bound increases at most linearly in the radius ε, uniformly for all probability measures P.

If the pairwise loss function is convex and separately continuous, then we show that the RPL operator $f_{L^\star,\mathrm{P},\lambda}$ has a bounded Gâteaux-derivative if a bounded and continuous kernel is used. As a special case we obtain a bounded influence function.

We also show that the sequence of RPL estimators $(f_{L^\star,\mathbb{P}_n,\lambda})_{n\in\mathbb{N}}$ is qualitatively robust for all probabiity measures P, if the pairwise loss function is convex, separately Lipschitz continuous with uniformly bounded partial derivatives up to order 2 and if the kernel is continuous and bounded.

Assume that all pairs of random variables $(X_i, Y_i) \overset{i.i.d.}{\sim} \mathrm{P}$ and denote the distribution of the RPL estimator $f_{L^\star,\mathbb{P}_n,\lambda}$ by $\mathscr{L}_n(S; \mathrm{P})$. Of course, this distribution

is unknown, because P is unknown in machine learning. To estimate this unknown probability measure, we use Efron's empirical bootstrap and replace P by the empirical distribution $\mathbb{P}_n$ to obtain $\mathscr{L}_n(S;\mathbb{P}_n)$ where the random variables $(X_i^*, Y_i^*) \overset{i.i.d.}{\sim} \mathbb{P}_n$. Note that $\mathscr{L}_n(S;\mathbb{P}_n)$ is a probability kernel, i.e. it can be considered as a *random* probability measure, but it can also be considered as a random variable in an abstract space. If $\mathcal{X} \times \mathcal{Y}$ is even a *compact* separable metric space, we show that then even the sequence $\big(\mathscr{L}_n(S;\mathbb{P}_n)\big)_{n\in\mathbb{N}}$ of bootstrap approximations is qualitatively robust for all probability measures P, provided the pairwise loss function L is convex, separately Lipschitz continuous with uniformly bounded partial derivatives up to order 2 and if the kernel is continuous and bounded.

The results are given in more detail in [2].

Acknowledgements. The work by A. Christmann described here is partially supported by a grant of the Deutsche Forschungsgesellschaft [Project No. CH/291/2-1]. The work by D.-X. Zhou described here is supported partially by a grant from the NSFC/RGC Joint Research Scheme [RGC Project No. N_CityU120/14 and NSFC Project No. 11461161006].

References

[1] Christmann, A., Salibían-Barrera, M., and Aelst, S. V., *Qualitative robustness of bootstrap approximations for kernel based methods*, In C. Becker, R. Fried, and S. Kuhnt, editors, Robustness and Complex Data Structures, pages 277–293. Springer, Heidelberg, 2013.

[2] Christmann, A. and Zhou, D. X., *On the Robustness of Regularized Pairwise Learning Methods Based on Kernels*, to appear in: J. of Complexity, 2016.

[3] Cucker, F. and Zhou, D. X., *Learning Theory: An Approximation Theory Viewpoint*, Cambridge University Press, Cambridge, 2007.

[4] Cuevas, A., *Qualitative robustness in abstract inference*, J. Statist. Plann. Inference, **18**, 277–289, 1988.

[5] Cuevas, A. and Romo, R., *On robustness properties of bootstrap approximations*, J. Statist. Plann. Inference, **37**, 181–191, 1993.

[6] Steinwart, I. and Christmann, *Support Vector Machines*. Springer, New York, 2008.

Analysis of sparse and low rank recovery via Mendelson's small ball method

Holger Rauhut

Compressive sensing [3] considers the recovery of (approximately) sparse vectors (signals, images etc.) from incomplete linear measurements via efficient algorithms. An extension of this theory replaces the sparsity assumption by a low rank assumption of a matrix to be recovered [11]. All provably optimal constructions of measurement matrices (modeling the process of taking measurements) known so far involve randomness. A recent method for estimating minima of certain stochastic processes from below due to Mendelson [10, 7] allows to significantly relax assumptions on the distribution of the measurement matrices and to study scenarios that were previously inaccessible with other methods.

In mathematical terms, we aim at reconstructing $x \in \mathbb{R}^N$ from

$$y = Ax, \quad A \in \mathbb{R}^{m\times N},$$

where $m \ll N$. Without further assumptions reconstruction is apparently impossible and compressive sensing supposes that x is sparse, i.e.,

$$\|x\|_0 = \#\{\ell : x_\ell \neq 0\} \leq s$$

for some $s < m \ll N$, or that it can at least be well-approximated by a sparse vector. Similarly, the low rank matrix recovery problem consists in reconstructing a matrix $X \in \mathbb{R}^{n_1 \times n_2}$ of rank at most r from

$$y = \mathcal{A}(X),$$

where $\mathcal{A} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^m$ is a linear map with $m \ll n_1 n_2$. While the recovery approaches of ℓ_0-minimization and rank-minimization are NP-hard, tractable algorithms for the recovery have been developed, most notably ℓ_1-minimization and nuclear norm minimization which consist in finding the minimizer of

$$\min_{z:Az=y} \|z\|_1 \qquad \text{and, respectively,} \qquad \min_{Z:\mathcal{A}(Z)=y} \|Z\|_*,$$

where $\|z\|_1 = \sum_{\ell=1}^N |z_\ell|$ and $\|Z\|_* = \sum_{j=1}^{\min\{n_1,n_2\}} \sigma_j(Z)$ is the nuclear norm with $\sigma_j(Z)$ being the singular values of Z. Standard results [3] state that for a random draw $A \in \mathbb{R}^{m \times N}$ of a Gaussian matrix, i.e., the entries of A being independent mean-zero, variance one, Gaussian variables, ℓ_1-minimization is able to reconstruct s-sparse vectors x exactly from $y = Ax$ with high probability provided

$$m \geq Cs \log(eN/s). \tag{1}$$

Similarly, nuclear norm minimization reconstructs rank-r matrices X from $y = \mathcal{A}(X)$ with $\mathcal{A}$ being a Gaussian random map with high probability provided $m \geq Cr(n_1 + n_2)$, see e.g. [11, 1].

Versions of the so-called null space property (NSP) characterize sparse and low rank recovery [3]. For $1 \leq p \leq \infty$ the ℓ_p-robust null space property of order s with constants $\rho \in (0,1)$ and $\tau > 0$ requires that

$$\|v_S\|_2 \leq \frac{\rho}{\sqrt{s}} \|v_{S^c}\|_1 + \tau \|Av\|_p \quad \text{for all } v \in \mathbb{R}^N, S \subset \{1,\dots,N\}, \#S \leq s, \tag{2}$$

where $S^c = \{1,2,\dots,N\} \setminus S$ denotes the complement of S, and v_S is the restriction of v to S. In fact, if this property holds, then for noisy measurements $y = Ax + e$ with $\|e\|_p \leq \eta$, the minimizer $x^\sharp$ of $\|z\|_1$ subject to $\|Az - y\|_p \leq \eta$ satisfies

$$\|x - x^\sharp\|_2 \leq C \frac{\sigma_s(x)_1}{\sqrt{s}} + D\tau\eta,$$

where $\sigma_s(x)_1 = \min\{\|x - z\|_1 : \|z\|_0 \leq s\}$ denotes the error of best s-term approximation in ℓ_1. While for $p = 2$, the ℓ_2-robust null space property is implied by the well-known restricted isometry property (RIP) [3], checking the version for $p \neq 2$ via variants of the RIP leads to highly suboptimal bounds on the required number of measurements, see e.g. the discussion in [2]. Therefore, a direct analysis of the NSP is conducted in [2], which also allows to relax assumptions on the distribution of the entries of the random matrix. Introducing the set

$$T_{\rho,s} = \{v \in \mathbf{R}^N : \|v_S\|_2 \geq \frac{\rho}{\sqrt{s}} \|v_{S^c}\|_1 \quad \text{for some } S \text{ with } \#S \leq s\},$$

the NSP (2) is ensured by the condition [6, 2]

$$\inf_{x \in T_{\rho,S}, \|x\|_2=1} \|Ax\|_p \geq \tau^{-1}. \tag{3}$$

This reformulation allows an analysis by Mendelson's small ball method [10, 7] (first introduced in the context of learning theory), which provides a general tool for bounding infima over quantities like $\|Ax\|_p$ for random A under rather weak assumptions.

Carrying through this analysis [2, 9] essentially shows that a random matrix with independent mean zero, variance one, and $\log(N)$-finite moments is able to recover (approximately) s-sparse vectors from noisy measurements with ℓ_p-bounded noise via ℓ_p-constrained ℓ_1-minimization with the optimal number (1) of measurements.

A similar condition like (3) implies stable and robust recovery of rank r matrices X from $y = \mathcal{A}(X)$ via nuclear norm minimization. Using the small ball method, it is shown in [5] that random measurement maps with independent, mean-zero, variance one and four finite moments, successfully recover rank r matrices provided that $m \geq Cr(n_1 + n_2)$. Robustness under ℓ_p-bounded noise on the measurements holds as well.

For the application of low rank matrix recovery in quantum tomography [4, 8], it is of interest to study the recovery from rank-one projections generated from randomly chosen elements of a so-called complex projective t-design. Exploiting again Mendelson's small ball method, it is shown in [8, 5] that $m \geq Crn\log(n)$ random measurements of an Hermitian rank r matrix $X \in \mathbb{C}^{n\times n}$, generated by an (approximate) 4-design, are sufficient for robust recovery with high probability.

References

[1] E. Candès, Y. Plan, Y, Tight oracle bounds for low-rank matrix recovery from a minimal number of random measurements. IEEE Trans. Inform. Theory, 57(4):2342-2359, 2011.

[2] S. Dirksen, G. Lecu'e, and H. Rauhut. On the gap between restricted isometry properties and sparse recovery conditions. *IEEE Trans. Inform. Theory*, to appear. DOI:10.1109/TIT.2016.2570244

[3] S. Foucart, H. Rauhut. *A Mathematical Introduction to Compressive Sensing*. Applied and Numerical Harmonic Analysis. Birkhäuser, 2013.

[4] D. Gross, Y. Liu, T. Flammia, S. Becker, J. Eisert Quantum state tomography via compressed sensing. Phys. Rev. Lett., 105, p. 150401, 2010.

[5] M. Kabanava, R. Kueng, H. Rauhut, U. Terstiege. Stable low rank matrix recovery via null space properties. Information and Inference, to appear. DOI:10.1093/imaiai/iaw014

[6] M. Kabanava and H. Rauhut. Analysis ℓ_1-recovery with frames and Gaussian measurements. *Acta Appl. Math.*, 140(1):173–195, 2015.

[7] V. Koltchinskii and S. Mendelson. Bounding the smallest singular value of a random matrix without concentration. *Internat. Math. Res. Notices*, 2015.

[8] R. Kueng, H. Rauhut, and U. Terstiege. Low rank matrix recovery from rank one measurements. *Appl. Comput. Harmon. Anal.*, to appear. DOI:10.1016/j.acha.2015.07.007

[9] G. Lecué and S. Mendelson. Sparse recovery under weak moment assumptions. *J. Eur. Math. Soc. (JEMS)*, to appear.

[10] S. Mendelson. Learning without Concentration. *J. ACM*, 62(3):1–25, 2015.

[11] B. Recht, M. Fazel, P. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. SIAM Rev., 52(3):471-501, 2010.

Nonnegative subdivision revisited

Kurt Jetter

It is well-known that uniform convergence of stationary d-variate subdivision with finite mask can be analyzed through left convergence of products of certain matrices from a finite alphabet. In case $\mathbf{a} = (\mathbf{a}(\alpha))_{\alpha\in\mathbb{Z}^d}$ is the mask – assumed to be finitely supported, w.l.o.g., in a cube

$$\text{supp } \mathbf{a} =: I \subseteq R_{\mathbf{N}} = \{0,1,\dots,N_1\} \times \{0,1,\dots,N_2\} \times \cdots \times \{0,1,\dots,N_d\} \subset \mathbb{Z}^d$$

where $N_1,\dots,N_d$ are positive integers – the subdivision operator is given by

$$\mathbf{S_a} \ : \quad \ell^\infty(\mathbb{Z}^d) \to \ell^\infty(\mathbb{Z}^d) \ , \quad \mathbf{c} \mapsto \mathbf{S_a}\,\mathbf{c} \ ,$$

with

$$(\mathbf{S_a}\,\mathbf{c})(\alpha) := \sum_{\beta\in\mathbb{Z}^d} \mathbf{a}(\alpha - 2\beta)\ \mathbf{c}(\beta) \ , \quad \alpha \in \mathbb{Z}^d \ .$$

Upon iteration we get the (stationary) subdivision scheme

$$\mathbf{c}^{(k+1)} := \mathbf{S_a}\ \mathbf{c}^{(k)} \ , \quad k = 0,1,\cdots \ ,$$

initialized with the starting sequence $\mathbf{c} = \mathbf{c}^{(0)} \in \ell^\infty(\mathbb{Z}^d)$. Uniform convergence of the scheme refers to the following property of the so-called iterated masks

$$\mathbf{a}^{(1)} = \mathbf{a} \ , \quad \mathbf{a}^{(k+1)} = \mathbf{S_a}\ \mathbf{a}^{(k)} \ , \ k \geq 1 \ ,$$

viz., the existence of a continuous d-variate function $\phi \neq 0$ such that

$$\lim_{k\to\infty} \sup_{\alpha\in\mathbb{Z}^d} \left|\phi\left(\frac{\alpha}{2^k}\right) - \mathbf{a}^{(k)}(\alpha)\right| = 0 \ .$$

In terms of the matrix $\mathbf{A} = \big(\mathbf{a}(-\alpha+2\beta)\big)_{\alpha,\beta\in R_{\mathbf{N}}}$, and of the system of submatrices (our alphabet)

$$\mathbf{A}_\delta = \big(\mathbf{a}(-(\alpha+\delta)+2(\beta+\delta)\big)_{\alpha,\beta\in\Gamma} \ , \quad \delta \in E^d := \{0,1\}^d \ ,$$

where

$$\Gamma = R_{\mathbf{N-1}} = \{0,1,\dots,N_1-1\} \times \{0,1,\dots,N_2-1\} \times \cdots \times \{0,1,\dots,N_d-1\} \ ,$$

we look at words from this alphabet of length k, i.e., products of k factors from our alphabet. Since for given $\delta_1,\delta_2,\dots,\delta_k \in E^d$ and $\lambda := \delta_1 + 2\delta_2 + \cdots + 2^{k-1}\delta_k$ we have

$$\mathbf{A}_{\delta_1}\mathbf{A}_{\delta_2}\cdots\mathbf{A}_{\delta_k}(\alpha,\beta) = a^{(k)}(-\alpha+\lambda+2^k\beta) \ , \quad \alpha,\beta \in R_{\mathbf{N-1}} \ ,$$

we can relate all non-zero coefficients of the iterated masks to the entries of finite words from our alphabet, and uniform convergence means

$$\lim_{k\to\infty} \left|\phi\left(-\frac{\alpha}{2^k} + \sum_{i=1}^{k}\frac{\epsilon_i}{2^i} + \beta\right) - \mathbf{A}_{\epsilon_k}\mathbf{A}_{\epsilon_{k-1}}\cdots\mathbf{A}_{\epsilon_1}(\alpha,\beta)\right| = 0 \quad \text{for} \quad \alpha,\beta \in R_{\mathbf{N-1}} \ .$$

In the limit, as $k \to \infty$, we have

$$-\frac{\alpha}{2^k} + \sum_{i=1}^{k} \frac{\epsilon_i}{2^i} \to \sum_{i=1}^{\infty} \frac{\epsilon_i}{2^i} =: \mathbf{x} \in [0,1]^d \quad \text{for} \quad \alpha \in R_{\mathbf{N-1}} \ .$$

Thus, uniform convergence means that the matrix product $\mathbf{A}_{\epsilon_k}\mathbf{A}_{\epsilon_{k-1}} \cdots \mathbf{A}_{\epsilon_1}$, for a given sequence $\epsilon_1, \epsilon_2, \cdots$ of 'multi-bits' from E, is left convergent to a rank-one matrix with equal rows of type

$$(\cdots \phi(\mathbf{x} + \beta) \cdots)_{\beta \in R_{\mathbf{N-1}}} \ ,$$

where $\mathbf{x} \in [0,1]^d$ has the dyadic expansion $\mathbf{x} = \sum_{i=1}^{\infty} \frac{\epsilon_i}{2^i}$.

In the case of nonnegative subdivision, where the mask and hence our alphabet of matrices is nonnegative, the necessary condition for convergence (the so-called sum rule) tells that the matrices $\mathbf{A}_\delta$ must be row stochastic. Therefore, the convergence problem can be described in terms of a nonhomogeneous finite Markov chain with the system $\{\mathbf{A}_\delta \ : \ \delta \in E^d\}$ as transition matrices. Convergence of such processes has been intensively studied in the past, and in our talk we refer to the convergence result in [1] which is based on Wolfowitz' [7] notion of SIA matrices, Hajnal's [3] notion of scrambling power

$$\gamma(\mathbf{P}) := \min_{i_1, i_2} \sum_j \min\{ \ p_{i_1,j}, p_{i_2,j} \ \},$$

for any square row stochastic matrix $\mathbf{P}$, and the notion of the corresponding ergodic coefficient

$$\tau(\mathbf{P}) := 1 - \gamma(\mathbf{P}) \ .$$

In this way, one can see that uniform convergence of nonnegative subdivision is equivalent to the fact that each word from our alphabet is SIA, and equivalently, each word of sufficient length must have the scrambling property (its ergodic coefficient must be less than one), or even must have a strictly positive column. The latter two properties refer to sign patterns of row stochastic matrices, and sufficient length just means that the length is at most equal to the number of possible sign patterns of such matrices.

An account of this approach to convergence of nonnegative subdivision refering to an nonhomogeneous Markov process is given in [4], where also a characterization in terms of directed graphs is presented. There also a slight extension of the Ren and Beard [6] partial characterization of the SIA property in terms of a tree property of the graph can be found.

Recent research on convergence of nonnegative subdivision has focused on properties of the support of the mask which can be checked in a faster way than checking all possible patterns for words. In univariate subdivision, the convergence problem was finally characterized by Xinlong Zhou, cf. [8, 9], who has shown that uniform convergence is equivalent to the greatest common divisor property for the indices from the support of the mask, subject that the sum rule being satisfied. A similar simple characterization in the multivariate case is still far from being available,

although some progress can be seen in work of Neumann [5], and the recent work of Li Cheng [2].

References

[1] J. M. Anthonisse and H. Tijms, *Exponential convergence of products of stochastic matrices*, J. Math. Anal. Appl. **59** (1977), 360–364.

[2] Li Cheng, *Combinatorial properties of multivariate subdivision schemes with nonnegative masks*, Dissertation, Universität Duisburg-Essen (2016).

[3] J. Hajnal, *Weak ergodicity in nonhomogeneous Markov chains*, Proc. Cambridge Philos. Soc. **54** (1958), 233–246.

[4] K. Jetter and X. Li, *SIA matrices and non-negative subdivision*, Results in Math. **62** (2012), 355–375.

[5] M. Neumann, *Konvergenz bivariater Subdivisions-Algorithmen mit nicht-negativen Masken*, Masterarbeit, Universität Duisburg-Essen (2013).

[6] W. Ren and R. W. Beard, *Consensus seeking in multi agent systems under dynamically changing interaction topologies*, IEEE Trans. on Automatic Control **50** (2005), 655–661.

[7] J. Wolfowitz, *Products of indecomposable, aperiodic, stochastic matrices*, Proc. Amer. Math. Soc. **14** (1963), 733–737.

[8] X.-L. Zhou, *Subdivision scheme with nonnegative masks*, Math. Comp. **47** (2005), 819–839.

[9] X.-L. Zhou, *Positivity of refinable functions defined by nonnegative masks*, Appl. Comput. Harmonic Analysis **27** (2009), 133–156.

Recovery of sparse exponential sums and sparse polynomials in several variables

Tomas Sauer

The goal is to reconstruct exponential functions of the form

$$f(x) = \sum_{\omega\in\Omega} f_\omega\, e^{\omega^T x}, \qquad \Omega \subset \mathbb{R}^s + i\,\mathbb{T}^s, \qquad \mathbb{T} := \mathbb{R}/2\pi\mathbb{Z}, \tag{1}$$

or polynomials

$$f(x) = \sum_{\alpha\in A} f_\alpha\, x^\alpha, \qquad A \subset \mathbb{N}_0^s, \tag{2}$$

from finitely many *samples* on the multiinteger grid $\mathbb{Z}^s$. The main assumption is that Ω and A are *sparse* sets, i.e., that they are of small cardinality, and that all the coefficients are different from zero so that the sums are "efficient". There is, however, no restriction on the size of the frequencies in (1) or the degree of the polynomial in (2). It is easy to see that (2) can be reduced to (1): if recovery of sparse exponentials from integer samples is possible, just choose an nonsingular matrix $\Xi \in \mathbb{C}^{s\times s}$ and consider $f(e^{\Xi\cdot})$ which recovers the coefficients f_α and the frequencies $\omega_\alpha = \Xi^T\alpha$, $\alpha \in A$, from which A can be be obtained by setting $\alpha = \operatorname{rd}\left(\Xi^{-T}\omega_\alpha\right)$, where the rounding even has a stabilizing effect and the frequencies can be computed by fast numerical methods as in [5], in contrast to applying univariate Prony to symbolic polynomials like in [2].

Recovery of sparse exponentials as in (1) is called *Prony's problem* and was stated and solved (at least in principle) as early as 1795 in [3]. The case of

several variables, however, attracted attention only very recently, for example in the context of *"superresolution"*. Like in the univariate case, the solution is based on considering *Hankel matrices* of the type

$$F_{A,B} = \left[f(\alpha+\beta) : \begin{array}{c} \alpha \in A \\ \beta \in B \end{array} \right] \in \mathbb{C}^{A\times B}, \qquad A, B \subset \mathbb{N}_0^s.$$

Defining $X_\Omega := \{e^\omega : \omega \in \Omega\} \subset \mathbb{C}^s$ and the *monomial Vandermonde matrix* $V(X,A) = \left[x^\alpha : \begin{array}{c} x \in X \\ \alpha \in A \end{array} \right]$, $X \subset \mathbb{C}^s$, $A \subset \mathbb{N}_0^s$, we get the two fundamental identities

$$F_{A,B} = V(X_\Omega, A)^T \, F_\Omega \, V(X_\Omega, B) \tag{3}$$

$$F_{A,B}\, p = V(X_\Omega, A)^T \, F_\Omega \, p(X_\Omega), \tag{4}$$

where $F_\Omega = \operatorname{diag}[f_\omega : \omega \in \Omega]$ and we identify a polynomial

$$p(x) = \sum_{\beta\in B} p_\beta \, x^\beta$$

with its coefficient vector $[p_\beta : \beta \in B] \in \mathbb{C}^s$. These factorizations quite readily yield information on the possibility to reconstruct Ω from measurements:

(1) the diagonal coefficient matrix F_Ω can be reconstructed from $F_{A,B}$ if and only if both Vandermonde matrices in (3) have rank $\geq \#\Omega$.
(2) if the rank of $V(X_\Omega, A)$ exceeds $\#\Omega$, then the rank of F_{A,Γ_n}, $\Gamma_n := \{\alpha \in \mathbb{N}_0^s : |\alpha| = n\}$, is the *affine Hilbert function* of the ideal $I_\Omega = \{p \in \mathbb{C}[x_1, \ldots, x_S] : p(X_\Omega) = 0\}$.

Taking into account these two observations, good choices for A are index sets such that $\Pi_A = \operatorname{span}\{(\cdot)^\alpha : \alpha \in A\}$ is a *universal interpolation space* of order $\#\Omega$, i.e., a space that allows for (non unique!) interpolation at *any* subset of $\mathbb{C}^s$ of cardinality $\leq \#\Omega$. If one request (total) *degree reduction* of the respective interpolation operator in addition, then the minimal index set A of order N can be identified as the *hyperbolic cross*

$$\Upsilon_N := \left\{ \alpha \in \mathbb{N}_0^s : \prod_{j=1}^s (\alpha_j + 1) \leq N \right\}.$$

Based on this knowledge, it is possible to compute a basis of the ideal I_Ω by successively adding columns of the matrix $F_{A,B}$ with a nested sequence of B which can be equipped with a simple termination condition based on the rank of $F_{A,B}$. Two canonical choices are to build B by adding simple monomials, leading to a symbolic method and a *Gröbner basis* for I_Ω, the second way is to update by blocks of total degree which results in a term order free *H–basis* that can be determined entirely by orthogonal projections and QR–decompositions from Numerical Linear Algebra which allows for computation in a floating point environment.

Once this *Prony ideal* is known, the frequencies can be determined by eigenvalue methods analogous to Frobenius companion matrices, cf. [1] and the coefficients by solving yet another Vandermonde system.

References

[1] W. Auzinger and H. J. Stetter, *An elimination algorithm for the computation of all zeros of a system of multivariate polynomial equations*, Numerical mathematics, Singapore 1988, Internat. Schriftenreihe Numer. Math., vol. 86, Birkhäuser, Basel, 1988, pp. 11–30.

[2] M. Ben-Or and P. Tiwari, *A deterministic algorithm for sparse multivariate polynomial interpolation*, Proc. Twentieth Annual ACM Symp. Theory Comput., ACM Press, New York, 1988, pp. 301–309.

[3] C. Prony, *Essai expérimental et analytique sur les lois de la dilabilité des fluides élastiques, et sur celles de la force expansive de la vapeur de l'eau et de la vapeur de l'alkool, à différentes températures*, J. de l'École polytechnique **2** (1795), 24–77.

[4] R. Roy and Th. Kailath, *ESPRIT – estimation of signal parameters via rotational invariance techniques*, IEEE Trans. Acoustics, Speech and Signal Processing **37** (1989), 984–995.

[5] T. Sauer, *Prony's method in several variables*, Submitted for publication (2015), arXiv:1602.02352.

Adaptive compression against countable alphabets

Stéphane Boucheron

(joint work with Anna Ben-Hamou, Elisabeth Gassiat)

We study the problem of lossless universal source coding for stationary memoryless sources on a countably infinite alphabet ($\mathcal{X} = \mathbb{N}$). Lossless compression consists of mapping messages (finite sequences of symbols from $\mathcal{X}$) to codewords (binary strings). The mapping has to be not only one-to-one but also uniquely decodable: any binary string should be parsed in at most one way into a sequence of codewords. The first aim of compression is to minimize the expected length of codewords.

If a single source (that is a probability distribution P over infinite sequences of symbols from $\mathcal{X}$) has to be handled, Shannon's first theorem asserts that the minimum expected length of codewords when encoding messages of length n is not smaller than the Shannon entropy of P^n (the trace of P over $\mathcal{X}^n$): $-\sum_{x\in\mathcal{X}^n} P^n(x)\log P^n(x)$. Up to an additive constant, this lower bound is achievable by well understood techniques. The most relevant to our work is arithmetic coding. This method takes advantage of a correspondence between uniquely decodable codes and probability distributions that is established thanks to the Kraft-McMillan inequality [11]. Indeed, if a uniquely decodable code is identified with a probability distribution Q^n over $\mathcal{X}^n$, the length of the codeword associated with $x \in \mathcal{X}^n$, is not larger than $1 - \log Q^n(x)$. The performance of a probability Q^n (henceforth called a coding probability) with respect to a sampling probability P^n is quantified by the redundancy, that is the expected difference between the ideal codeword length $-\log P^n(x)$ and the actual codeword length:

$$\sum_{x\in\mathcal{X}^n} P^n(x)\log\frac{P^n(x)}{Q^n(x)}$$

which is also the relative entropy $D(P^n \mid Q^n)$ between the sampling probability and the coding probability.

The problem of universal coding arises when handling a collection $\mathcal{C}$ of sources. Universal coding aims at building a single probability distribution Q^n that performs well with respect to the whole class. The performance is measured by worst case redundancy

$$\overline{R}(Q^n, \mathcal{C}^n) := \max_{P \in \mathcal{C}} D(P^n \mid Q^n) \,.$$

The minimax redundancy

$$\overline{R}(\mathcal{C}^n) := \min_{Q^n} \max_{P \in \mathcal{C}} D(P^n \mid Q^n)$$

quantifies the difficulty of universal coding with respect to class $\mathcal{C}^n$. When the alphabet is finite, universal coding is well understood and may be spectacularly successful: if $\mathcal{C}$ consists of all stationary memoryless sources over an alphabet of size k, then

$$\overline{R}(\mathcal{C}^n) = \frac{k-1}{2} \log \frac{n}{2\pi \mathrm{e}} + O_k(1) \,,$$

see [3, 18, 19] and references therein.

If the considered collection of sources is too large, minimax redundancy may turn out to be trivial (scale linearly with message length). In other settings, the source class may be the union of smaller classes with widely differing minimax redundancy rates (for example sources defined by finite context trees over a finite alphabet have redundancy rates that depend on the shape of the context tree). *Adaptive coding* then considers an appropriate, more general setting. Assume that the excessively large collection of sources is the union of smaller subclasses and that, for each subclass, minimax redundancy rate is non trivial and a good universal coder is available. Is it then possible to engineer a single coding method that performs well over all subclasses in the collection? This problem is related to competitive estimation, it could be called competitive coding. Adaptive coding is also known as twice-universal coding. We conform to the conventions and definitions of mathematical statistics. Context-Tree-Weighting [2] provides an example of an adaptive code with respect to sources with bounded or unbounded memory over finite alphabets.

Let $(\mathcal{C}(\alpha))$ be a collection of source classes indexed by $\alpha \in A$. A sequence $(Q^n)_{n \geq 1}$ of coding probabilities is said to be *asymptotically adaptive* with respect to a collection $(\mathcal{C}(\alpha))_{\alpha \in A}$ of source classes if for all $\alpha \in A$

$$\overline{R}(Q^n, \mathcal{C}(\alpha)^n) = \sup_{P \in \mathcal{C}(\alpha)} D(P^n, Q^n) \leq (1 + o_\alpha(1)) \overline{R}(\mathcal{C}(\alpha)^n) \tag{1}$$

as n tends to infinity. If the inequality (1) holds with a factor other than $(1+o_\alpha(1))$ (that may depend on α) larger than 1 to the right, then we say that there is adaptivity *within* this factor. Note that Q_n cannot depend on α or else the problem is simply one of universality.

When the alphabet is countably infinite, even if we focus on stationary memoryless sources, universal coding is not achievable [14]. Several competing approaches to this problem have been considered.

(1) [15] separate the description of strings over large alphabets into two parts: description of the symbols appearing in the string, and of their pattern, the order in which the symbols appear. They redefine the performance criterion by focusing on compressing the message's *pattern* [15];
(2) investigating the redundancy on smaller source classes that satisfy Kieffer's condition. The so-called envelope classes investigated in [8] form an example of such classes [1, 17].

In pattern coding, each symbol is replaced by the rank of its first occurrence. For instance, the pattern of the message *abracadabra* is

$$12314151231\,.$$

By setting aside the actual value of symbols, pattern coding focuses on the structure of messages. When the source is stationary and memoryless, all the relevant information on the pattern is contained in the *profile* of the sample, recording the number of symbols occurring once, twice... [16, 15] have shown (among other things) that a variant of Shtarkov's normalized maximum likelihood (NML) coder achieves a non-trivial redundancy of the collection of pattern distributions induced by stationary memoryless sources over a countable alphabet.

This paper pursues both lines of research: we deal with collection of so-called envelope classes, but the adaptive code we introduce and investigate will turn out to be a pattern encoder. In contrast with [15], we attempt to handle both dictionary and pattern encoding, that is to interleave dictionary encoding and pattern coding.

Let f be a non-increasing mapping from $\mathbb{N}_* := \mathbb{N} \setminus \{0\}$ to $(0,1]$, with $1 < \sum_{j\in\mathbb{N}_*} f(j) < \infty$. The *envelope class* $\mathcal{C}(f)$ defined by the function f is the collection of distributions which are dominated by f: $\mathcal{C}(f) := \left\{P : \forall j \in \mathbb{N}_*,\ p_j \leq f(j)\ \right\}$. Define $\ell_f = \min\left\{\ell \geq 1,\ \sum_{j=\ell}^{+\infty} f(j) \leq 1\right\}$. The associated *envelope distribution* F is defined as $F(k) := 1 - \sum_{j>k} f(j)$ if $k+1 \geq \ell_f$, and $F(k) := 0$ otherwise.

Envelope classes provide a framework where the search for adaptive coding strategies is feasible. In previous papers on adaptive coding against envelope classes, envelopes were defined by assuming some conditions on the decay of the envelope survival function $(1-F)$ or equivalently on the envelope quantile function. In this paper, where we are interested in the full range of *regularly varying* envelope classes, we find it convenient to introduce the unifying framework proposed by Karlin in [13]. This framework has recently received attention in random combinatorics and stochastic processes theory, see [12] for a survey.

A probability mass function $(p_j)_{j\geq 1}$ defines a *counting measure* ν defined by $\nu(\mathrm{d}x) = \sum_{j\geq 1} \delta_{p_j}(\mathrm{d}x)$, where δ_p is the Dirac mass at p. Let the counting function $\vec{\nu}(\cdot)$ be the right tail of ν, that is for all $x \in (0,1]$,

$$\vec{\nu}(x) = \nu[x,\infty[= |\{j \geq 1,\ p_j \geq x\}|\,.$$

In the text, we denote by ν_f, $\vec{\nu}_f$, $\nu_{1,f}$ the corresponding quantities when the underlying distribution is given by the envelope frequencies $(f_j)_{j\geq 1}$.

Karlin's setting proves illuminating. We revisit the tight redundancy bounds derived in [1]. If the envelope f satisfies a so-called regular variation condition ($\vec{\nu}(1/\cdot) \in \mathrm{RV}_\alpha, \alpha \in [0,1)$), the minimax redundancy rate $\overline{R}(\Lambda_f^n)$ scales like

$$\mathsf{R}_f(n) \quad := \quad \log(\mathrm{e}) \int_1^n \frac{\vec{\nu}_f(1/t)}{2t} \mathrm{d}t \,.$$

This characterization is a powerful generalization of the tight bounds that have been established for memoryless sources over finite alphabets. The latter can be regarded as envelope classes where $\vec{\nu}_f(x) = k$ for some k and all small enough x. Indeed $\frac{k-1}{2} \log n$ scales like $\log(\mathrm{e}) \int_1^n \frac{\vec{\nu}_f(1/t)}{2t} \mathrm{d}t$ with respect to both k and n. The integral of the counting function also provides an equivalent of the minimax redundancy for envelope classes defined by log-concave envelopes (such that $f(k)f(k+2)/f(k+1)^2 \leq 1$ for all $k \geq 1$) that was characterized in [7]. Up to a constant factor, the integral of the counting function also provides an equivalent of the minimax redundancy for envelope classes defined by envelopes with positive regular variation indexes as investigated in [10].

Revisiting the bounds on minimax redundancy rates from [1] using Karlin's setting also suggests a universal coding strategy for each envelope class. In words, when encoding the n^{th} symbol in the message, it seems sensible to handle symbols with probability larger than $1/n$ (frequent symbols) differently from symbols with probability smaller than $1/n$ (rare symbols). The probability of frequent symbols can be relatively faithfully estimated while the probability of rare symbols can barely be estimated from the message. The censoring code approach described in [8] explores that kind of path but cutoffs are built from conservative upper bounds on minimax redundancy rates. The adaptive censoring code approach described in [6, 7] implement this method in serendipitous way: with high probability, the sample maximum approximates the suggested cutoff.

In this paper, we combine pattern coding and censoring so as to manufacture a simple encoder that achieves adaptivity within a $\log\log n$ factor with respect to all envelope classes with regular variation index in $[0,1)$. This leads to the Pattern Censoring Code.

The main result is the next theorem.

Theorem 1. *Let $(Q_n)_n$ be the sequence of coding distributions associated with the Pattern Censoring Code. For all $\alpha \in [0,1[$, for every envelope function f with $\vec{\nu}_f(1/\cdot) \in \mathrm{RV}_\alpha$, there exists constants $a_f, b_f > 0$ such that*

$$(a_f + o_f(1)) \leq \frac{\overline{R}(\mathcal{C}^n(f))}{R_f(n)} \leq \frac{\overline{R}(Q_n, \mathcal{C}^n(f)}{R_f(n)} \leq (b_f + o_f(1)) \log\log n \,.$$

In particular, the Pattern Censoring Code is adaptive, within a $\log\log n$ factor, with respect to the collection

$$\big\{\mathcal{C}(f) : f \in \mathrm{RV}_\alpha, \alpha \in [0,1[\big\} \,.$$

References

[1] J. Acharya, A. Jafarpour, A. Orlitsky, and A. T. Suresh. Poissonization and universal compression of envelope classes, February 2014.

[2] O. Catoni. *Statistical learning theory and stochastic optimization*, volume 1851 of *Lecture Notes in Mathematics*. Springer-Verlag, 2004. Ecole d'Ete de Probabilites de Saint-Flour XXXI.

[3] B. Clarke and A. Barron. Jeffrey's prior is asymptotically least favorable under entropy risk. *J. Stat. Planning and Inference*, 41:37–60, 1994.

[4] A. Ben-Hamou, S. Boucheron, and M. I. Ohannessian. Concentration inequalities in the infinite urn scheme for occupancy counts and the missing mass, with applications. *Bernoulli*, to appear, 2016.

[5] N. Bingham, C. Goldie, and J. Teugels. *Regular variation*, volume 27. Cambridge University Press, 1989.

[6] D. Bontemps. Universal coding on infinite alphabets: exponentially decreasing envelopes. *IEEE Trans. Inform. Theory*, 57(3):1466–1478, 2011.

[7] D. Bontemps, S. Boucheron, and E. Gassiat. About adaptive coding on countable alphabets. *IEEE Trans. Inform. Theory*, 60(2):808–821, 2014.

[8] S. Boucheron, A. Garivier, and E. Gassiat. Coding over Infinite Alphabets. *IEEE Trans. Inform. Theory*, 55:358–373, 2009.

[9] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. OUP Oxford, 2013.

[10] S. Boucheron, E. Gassiat, and M. Ohannessian. About adaptive coding on countable alphabets: max-stable envelope classes. *IEEE Trans. Inform. Theory*, 61(9):4948–4967, 2015.

[11] T. Cover and J. Thomas. *Elements of information theory*. John Wiley & sons, 1991.

[12] A. Gnedin, B. Hansen, and J. Pitman. Notes on the occupancy problem with infinitely many boxes: general asymptotics and power laws. *Probab. Surv.*, 4:146–171, 2007.

[13] S. Karlin. Central limit theorems for certain infinite urn schemes. *J. Math. Mech.*, 17:373–401, 1967.

[14] J. C. Kieffer. A unified approach to weak universal source coding. *IEEE Trans. Inform. Theory*, 24(6):674–682, 1978.

[15] A. Orlitsky and N. P. Santhanam. Speaking of infinity. *IEEE Trans. Inform. Theory*, 50 (10):2215–2230, 2004.

[16] A. Orlitsky, N. P. Santhanam, and J. Zhang. Always good Turing: asymptotically optimal probability estimation. *Science*, 302(5644):427–431, 2003.

[17] X. Yang and A. Barron. Large Alphabet Coding and Prediction through Poissonization and Tilting. In *The Sixth Workshop on Information Theoretic Methods in Science and Engineering*, Tokyo, August 2013.

[18] Q. Xie and A. R. Barron. Minimax redundancy for the class of memoryless sources. *IEEE Trans. Inform. Theory*, 43:646–656, 1997.

[19] Q. Xie and A. R. Barron. Asymptotic minimax regret for data compression, gambling and prediction. *IEEE Trans. Inform. Theory*, 46:431–445, 2000.

Online learning with pairwise loss functions

Yiming Ying

Pairwise learning usually refers to a learning task which involves a loss function depending on pairs of instances. That is, the loss function depends on a pair of instances which, for any (x, y) and (x', y'), can be expressed by $\ell(f(x, x'), y, y')$ for a hypothesis function $f : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Most notable examples of pairwise learning include bipartite ranking [3], metric learning [5], minimum entropy error principle

[6] and AUC maximization [7]. Assume the data $\{z_i = (x_i, y_i) : i = 1, \ldots, T\}$ is i.i.d. drawn from an unknown distribution ρ on $\mathcal{X} \times \mathcal{Y}$. A unified formulation for such pairwise learning methods can be formulated as

$$\min_{f \in \mathcal{H}_K} \Big\{ \frac{1}{T(T-1)} \sum_{i,j=1, i \neq j}^{T} \ell(f(x_i, x_j), y_i, y_j) + \frac{\lambda}{2} \|f\|_K^2 \Big\}, \tag{1}$$

where $K : \mathcal{X}^2 \times \mathcal{X}^2 \to \mathbb{R}$ is a reproducing kernel and $\mathcal{H}_K$is the corresponding RKHS on $\mathcal{X}^2$ with norm $\|\cdot\|_K$. Statistical analysis for (1) in the batch learning setting was established in terms of algorithmic stability [1], robustness [2], and U-statistics [3].

Online learning algorithms are widely used in practice to deal with the large scale (streaming) data. However, most of such algorithms focused on the pointwise learning problems in classification and regression. There are a number of specific challenges in developing and analyzing online pairwise learning algorithms: 1) the objective function is usually defined over pairs of instances which is quadratic in the number of individual instances; 2) pairwise learning involves statistically dependent pairs of instances, which is fundamentally different from the i.i.d. assumption in classification and regression. Our main purpose is to develop online pairwise learning algorithms which are, in both algorithmic implementation and theoretical analysis, on a par with online algorithms in classification.

We start with a general online learning algorithm for pairwise learning in an unconstrained setting of a reproducing kernel Hilbert space (RKHS) which is given as follows:

Online Pairwise Learning Algorithm: Given $\lambda \geq 0$, initialize $f_1 = f_2 = 0$ and repeat, for any $2 \leq t \leq T$,

$$f_{t+1} = f_t - \gamma_t \Big[\frac{1}{t-1} \sum_{j=1}^{t-1} \ell'(f_t(x_t, x_j), y_t, y_j) K_{(x_t, x_j)} + \lambda f_t \Big], \tag{2}$$

where $\{\gamma_t > 0 : t \in \mathbb{N}\}$ is called step sizes.

We establish convergence analysis of the last iterate f_{T+1} for the above online algorithm in both regularized ($\lambda > 0$) and un-regularized ($\lambda = 0$) settings. In particular, for the regularized case with the hinge loss, we show, by properly choosing the step sizes, that the convergence rates are the same as the pointwise learning setting. For the un-regularized case, we focus on the least square loss and show that the excess generalization error is bounded by an K-functional in approximation theory plus a small error term. This convergence result shows that the properly chosen step sizes implicitly play the role of regularization.

The above general online algorithms require to store the previous instances $\{z_1, \ldots, z_t\}$ at iteration t which is not memory efficient. For the notable example of pairwise learning called AUC maximization, we can develop a truly online algorithm for which the space and per-iteration complexities only depend linearly on one datum. The key idea behind this is a novel formulation of AUC maximization as a stochastic saddle point problem (SPP). A stochastic online algorithm for

AUC maximization is then proposed and its convergence analysis is established. Experiments on various datasets show encouraging performance of the proposed algorithm.

The talk is based on the recent work [4, 8, 9] jointly with my collaborators: Prof. Ding-Xuan Zhou from City University of Hong Kong, Dr. Zheng-Chu Guo from Zhejiang University, Prof. Siwei Lyu and Dr. Longyin Wen from SUNY Albany.

References

[1] S. Agarwal and P. Niyogi. Generalization bounds for ranking algorithms via algorithmic stability. Journal of Machine Learning Research, **10** (2009), 441–474.

[2] A. Christmann and D. X. Zhou. On the robustness of regularized pairwise learning methods based on kernels. To appear in Journal of Complexity, 2016.

[3] S. Clemencon, G. Lugosi, and N. Vayatis. Ranking and empirical minimization of U-statistics. The Annals of Statistics, **36** (2008), 844-874.

[4] Z. C. Guo, Y. Ying and D. X. Zhou. Online regularized pairwise learning algorithms. Under minor revision for Advances in Computational Mathematics, 2016.

[5] K. Q. Weinberger and L. Saul. Distance metric learning for large margin nearest neighbour classification. Journal of Machine Learning Research, **10** (2009), 207–244.

[6] J. Fan, T. Hu, Q. Wu and D. X. Zhou. Consistency analysis of an empirical minimum error entropy algorithm. Applied and Computational Harmonic Analysis, **41** (2016), 164–189.

[7] P. Zhao, S. C. H. Hoi, R. Jin and T. Yang. Online AUC Maximization. In Proceedings of the 28th International Conference on Machine Learning (ICML), 2011.

[8] Y. Ying. S. Lyu and L. Wen. Stochastic online AUC maximization. Submitted for publication, 2016.

[9] Y. Ying and D. X. Zhou. Online pairwise leanring algorithms. Neural Computation, **28** (2016), 743–777.

Some learning algorithms for quantile regression

Dao-Hong Xiang

(joint work with Jia Cai, Ting Hu, and Ding-Xuan Zhou)

Quantile regression is a classical statistical method which results in estimates approximating either the median or other quantiles of the response variable. Compared with the least squares regression, quantile regression provides richer information about response variables such as stretching or compressing tails. We study three learning algorithms for quantile regression, which include: quantile regression with varying Gaussians, coefficient-based conditional quantile regression and learning with varying ϵ-insensitive pinball loss.

Allowing varying Gaussian kernels in the algorithms improves learning rates measured by regularization error and sample error. Special structures of Gaussian kernels enable us to construct, by a nice approximation scheme with a Fourier analysis technique, uniformly bounded regularizing functions achieving polynomial decays of the regularization error under a Sobolev smoothness condition. The sample error is estimated by using a projection operator and a tight bound for the

covering numbers of reproducing kernel Hilbert spaces generated by Gaussian kernels. Since pinball loss in the quantile regression setting has no strong convexity, we would not expect a variance-expectation bound for a general distribution. Hence, some kind of noise assumption on the distribution (see [1, 2]) plays an important role in our analysis. For more details, please see [4].

The original motivation of coefficient-based regularization schemes comes from the linear programming SVMs [3]. Among these algorithms, the hypothesis space is data dependent and the regularization term is the $\ell^q (1 \leq q \leq 2)$ norm of coefficients. In particular, the ℓ^1 norm plays an essential role in the Lasso algorithm in statistics and in the literature of compressed sensing, since ℓ^1-regularization usually leads to sparse solutions. In our error analysis, the main difficulty lies in the lack of a proper characterization of approximation error. We use a stepping stone technique to construct a function approximating the target function by means of integral operator. Fast learning rates are achieved in a general setting under mild conditions. For more details, please see [6].

We study the learning algorithms with varying ϵ-insensitive pinball loss which is motivated by the ϵ-insensitive loss for support vector regression and the pinball loss for quantile regression. The original motivation [3] for the insensitive parameter ϵ is that for balancing the approximation and sparsity of the algorithm, ϵ should change with the sample size. We solve the mathematical analysis for this original algorithms by studying a general learning algorithm. The learning rates are explicitly derived under a priori condition on approximation and capacity of the reproducing kernel Hilbert space. For more details, please see [5].

The work by D. H. Xiang is supported by the National Natural Science Foundation of China under Grant 11471292 and the Alexander von Humboldt Foundation of Germany.

References

[1] I. Steinwart, A. Christmann, *How SVMs can estimate quantiles and the median*, Adv. Neural Inf. Process. Syst. **20** (2008), 305–312.

[2] I. Steinwart, A. Christmann, *Estimating conditional quantiles with the help of the pinball loss*, Bernoulli **17** (2011), 211–225.

[3] V. Vapnik,*Statistical learning theory*, New York:Wiley, 1998.

[4] D. H. Xiang, *Conditional quantiles with varying Gaussians*, Adv. Comput. Math. **38**, (2013), 723–735.

[5] D. H. Xiang, T. Hu, D. X. Zhou, *Approximation analysis of learning algorithms for support vector regression and quantile regression*, Journal of Applied Mathematics **2012**, (2012), 17 pages, doi: 10.1155/2012/902139.

[6] J. Cai, D. H. Xiang, *Statistical consistency of coefficient-based conditional quantile regression*, Journal of Multivariate Analysis, **149**, (2016), 1–12. 723–735.

On the convergence of randomized Kaczmarz algorithm in Hilbert space

Xin Guo
(joint work with Junhong Lin, Ding-Xuan Zhou)

The classical Kaczmarz algorithm, designed to solve linear equation systems $Ax = y$ of finite dimension, was introduced by Kaczmarz in 1937. The algorithm iteratively uses the rows of the matrix A as measures, and in each step, projects the error vector onto the hyperplane perpendicular to the row vector selected. The convergence is well understood. It is observed that the speed of convergence strongly depends on the order how the rows of the matrix A are arranged. To remove this dependence, randomized Kaczmarz algorithm was introduced. In [Strohmer and Vershynin 2009], by defining the probabilities of the rows of the matrix A according to the squares of their norms, exponential convergence was obtained. For general distributions of measure vectors, exponential convergence was given by [Chen and Powell 2012]. Relaxed randomized Kaczmarz algorithm was introduced for noisy observations. In particular, [Needell 2010] shows that without the help of relaxation, randomized Kaczmarz algorithm will not converge with the presence of observational noise. Using the learning theory approaches, [Lin and Zhou 2015] gives a sufficient and necessary condition on the step sizes for the relaxed randomized Kaczmarz algorithm to converge, as well as an upper bound of the converging rate.

In the existing analysis of randomized Kaczmarz algorithms under the setting of no noise, most of the results give an exponential speed of decay of expected error. However, the base of such exponential rate is very close to one, with a small gap roughly proportional to the square of the smallest singular value of the coefficient matrix A. This problem restricts the application of the analysis in at least two aspects. First, for applications, especially when the coefficient matrix is large which is typically the scenario where the Kaczmarz algorithm is useful compared with traditional linear equation solvers, the smallest singular value of the coefficient matrix could be very close to zero, making the analysis not so useful. Second, it is impossible to generalize the analysis to Hilbert space, which covers many useful applications in learning theory and functional data analysis.

In the presented work, we develop the convergence analysis of the randomized Kaczmarz algorithm in Hilbert space. We show that the nature of the convergence is indeed a weak convergence with a polynomial rate. We give a concrete example which demonstrates that as long as there is observational noise with positive variance, no matter how small the variance is, the expected strong norm of the error vector can diverge to infinity if the step sizes are set to be a constant. On one hand, this shows that one should not, in general, expect randomized Kaczmarz algorithm to converge in the sense of expected norm. On the other hand, weak convergence is widely used in learning theory because it well corresponds to the strong convergence in the L^2 norm sense which is usually good enough for applications.

Real algebraic geometry for the construction of tight wavelet frames

Joachim Stöckler
(joint work with Maria Charina, Mihai Putinar, Claus Scheiderer)

We combine methods of real algebraic geometry, linear system theory and harmonic analysis for the construction and parameterization of classes of tight wavelet frames.

The construction of tight wavelet frames is often divided into two steps. First one picks a "nice" function $\phi \in L^2(\mathbb{R}^d)$, the *scaling function* of a multiresolution analysis, which has compact support, desirable smoothness and approximation properties and satisfies a refinement relation

$$\phi(x) = \sum_{k\in\mathbb{Z}^d} p_k \phi(2x-k)$$

with finitely many non-zero coefficients p_k. Secondly, one finds the framelets (or frame-generators)

$$\psi_\ell(x) = \sum_{k\in\mathbb{Z}^d} q_{\ell,k}\phi(2x-k), \qquad 1 \le \ell \le L,$$

by specifying their coefficient sequences $(q_{\ell,k})_{k\in\mathbb{Z}^d}$, such that the triple-indexed family

$$\Psi = \{2^{jd/2}\psi_\ell(2^j\cdot - k) \ : \ j\in\mathbb{Z},\ k\in\mathbb{Z}^d,\ \ell = 1,\dots,L\}$$

is a *tight frame* of $L^2(\mathbb{R}^d)$; this means that

$$\|f\|^2_{L^2(\mathbb{R}^d)} = \sum_{g\in\Psi} |\,\langle f,g\rangle\,|^2$$

holds for all functions $f \in L^2(\mathbb{R}^d)$. Tight wavelet frames were used by Zhang et al. [4] as multiscale kernels in Learning Theory.

Common criteria for obtaining suitable sequences $(q_{\ell,k})$ are specified in terms of the trigonometric polynomials $P(\xi) = \sum_k p_k e^{2\pi i k\xi}$ and Q_ℓ alike. Leaving out some technical details, which are not essential for this presentation, we arrive at the Unitary Extension Principle (UEP) which finds Q_ℓ by solving a matrix factorization problem

$$I - F(\xi)F(\xi)^* = G(\xi)G(\xi)^*. \tag{1}$$

Here, F is a column vector of trigonometric polynomials, which is easily obtained from P and such that $I - FF^*$ is positive semi-definite for all $\xi \in \mathbb{R}^d$. A more comprehensive approach to tight wavelet frames is the Oblique Extension Principle (OEP) which is based on the matrix factorization problem

$$K(\xi) - F(\xi)L(\xi)F(\xi)^* = G(\xi)G(\xi)^*. \tag{2}$$

Here, an additional positive definite matrix K and positive scalar L are chosen based on "vanishing moment properties" of the desired framelets.

Both factorization problems (1) and (2) are related to Hilbert's 17th problem in real algebraic geometry. Whereas solutions for the 1-dimensional case ($d = 1$) are

easily obtained from the Riesz-Fejer lemma, the situation is much more difficult for higher dimensions. Existence of solutions for $d = 2$ follows from a much more general result of C. Scheiderer [3], non-existence of trigonometric polynomial matrices G in (1) for $d = 3$ and for particular choices of P was shown in [1]. The results of our recent work can be summarized as follows:

1. The matrix factorization in (1) exists with trigonometric polynomial matrix G if and only if the "adjoint" scalar factorization
$$1 - F(\xi)^* F(\xi) = H(\xi)^* H(\xi)$$
exists with a vector H of trigonometric polynomials. Passing from H to G is constructive, and was already described in [2].
2. The matrix factorization in (2) exists with rational trigonometric matrix G if and only if the "adjoint" scalar factorization
$$\frac{1}{L(\xi)} - F(\xi)^* K(\xi)^{-1} F(\xi) = H(\xi)^* H(\xi)$$
exists with a vector H of rational trigonometric functions. Passing from H to G is constructive.

Regardless of the dimension, the factorization in (1) can be explicitly specified if ϕ is a multivariate box-spline.

A complementary method for constructing the matrix G in (1) was described in [1]. It defines F to be a complex polynomial vector on the polydisc $\mathbb{D}^d$. As part of the quite recent theory of multidimensional linear systems, the factorization with a polynomial matrix G in (1) exists if and only if F is an element of the Schur-Agler class of holomorphic functions. For $d = 1$ and $d = 2$, there are known algorithms for finding the representation of F as the transfer function of a linear system,
$$F(z) = A + BZ(I - DZ)^{-1}C,$$
where $Z = z_1 I_{n_1} \oplus \cdots \oplus z_d I_{n_d}$ is a diagonal matrix of monomials $z_1, \ldots, z_d$, A, B, C, D are complex matrices and
$$\begin{pmatrix} A & B \\ C & D \end{pmatrix} \quad \text{is a contraction.}$$

This connection to linear system theory was used in [1] in order to improve all previously known constructions of a tight wavelet frame of linear 2-dimensional box-splines as it achieves the smallest number of generators $L = 5$. The extension of this method to OEP constructions is under current investigation.

References

[1] M. Charina, M. Putinar, C. Scheiderer, J. Stöckler, *An algebraic perspective on multivariate tight wavelet frames*, Constr. Approx. **38** (2013), 253–276.

[2] M. J. Lai, J. Stöckler, *Construction of multivariate compactly supported tight wavelet frames*, Appl. Comp. Harmonic Anal. **21** (2006), 324–348.

[3] C. Scheiderer, *Sums of squares on real algebraic surfaces*, Manuscripta Math. **119** (2006), 395–410.

[4] W.-F. Zhang, D.-Q. Dai, H. Yan, *Framelet kernels with applications to SVR and RN*, IEEE Trans. on Systems, Man, and Cybernetics-Part B, **40** (2010), 1128–1144.

System theory: Learning orthogonal multi-wavelets

Maria Charina
(joint work with Costanza Conti, Mariantonia Cotronei)

One of the classical problems of approximation theory studies the existence and approximation properties of compactly supported wavelets and vector- or matrix-valued orthogonal multi-wavelets. Usually, wavelets or multi-wavelets are generated by a function $\psi : \mathbb{R} \to \mathbb{C}^{n\times m}$, $n \le m$, which is a finite linear combination

$$\psi = \sum_{k=0}^{s} \phi(2\cdot -k)q_k, \quad q_k \in \mathbb{C}^{m\times m},$$

of scaled integer shifts of the corresponding refinable function

$$\phi : \mathbb{R} \to \mathbb{C}^{n\times m}, \quad \phi = \sum_{k=0}^{s} \phi(2\cdot -k)p_k, \quad p_k \in \mathbb{C}^{m\times m}.$$

The mathematical challenge is to determine the classes of all masks $p = \{p_k\}_{k=0}^s$ and $q = \{q_k\}_{k=0}^s$ that guarantee the existence of such square-integrable ϕ and $\psi = (\psi_1, \dots, \psi_m)$ for which the set

$$\{2^{j/2}\psi_\ell(2^j\cdot -k) \ : \ j,k \in \mathbb{Z}, \ \ell = 1,\dots,m\}$$

is an orthonormal basis of $L_2^n(\mathbb{R})$. In the case $n = m = 1$, the elegant construction by Daubechies [3] yields such masks p and q with optimal approximation property: polynomials of degree less or equal to $s-2$ are in the span of the integer shifts of ϕ. For all other constellations of $n \le m$, the problem has resisted to be fully understood for 30+ years.

We show that there is no conceptual difference between wavelet ($n = m = 1$) and multi-wavelet constructions and provide their complete and unifying characterization. This characterization is based on classical results from system theory.

The link between wavelet and multi-wavelet constructions and system theory is offered by the so-called Unitary Extension Principle [5]. In contrast to [2], we do not assume that p is given and our goal is to construct all appropriate masks p and q simultaneously. The Unitary Extension Principle requires that the matrix polynomial

$$F : \mathbb{C} \to \mathbb{C}^{2m\times 2m}, \quad F(z) = \sum_{k=0}^{d} F_k z^k, \quad d = \lfloor s/2 \rfloor,$$

with matrix coefficients

$$F_k = \begin{pmatrix} p_{2k} & q_{2k} \\ p_{2k+1} & q_{2k+1} \end{pmatrix}, \quad k = 0,\dots,d,$$

is unitary on the unit circle. The unitarity of F and the maximum principle imply that $F(z)$ is contractive for any $|z| < 1$. Such holomorphic functions F, by [1, 6], belong to the Schur-Agler class, i.e. they are of the form

$$F(z) = A + Bz(I - Dz)^{-1} C, \quad |z| < 1,$$

with the unitary block matrix

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix} : \begin{matrix} \mathbb{C}^{2m} \\ \oplus \\ \mathbb{C}^{2md} \end{matrix} \rightarrow \begin{matrix} \mathbb{C}^{2m} \\ \oplus \\ \mathbb{C}^{2md}. \end{matrix}$$

Our main result characterizes the structure of all appropriate $F(z)$. We show that the corresponding $ABCD$-matrix is the product

$$\left(\begin{array}{c|cccc} F_0 & F_d & \dots & \dots & F_1 \\ \hline F_1 & F_0 & F_d & & \vdots \\ \vdots & F_1 & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & F_d \\ F_d & F_{d-1} & \dots & F_1 & F_0 \end{array}\right) \left(\begin{array}{c|c} I & 0 \\ \hline 0 & U \end{array}\right)$$

of a block circulant matrix and a unitary matrix with

$$U = \begin{pmatrix} F_0^* + F_d^* & F_1^* & \dots & F_{d-1}^* \\ F_{d-1}^* & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & F_1^* \\ F_1^* & \dots & F_{d-1}^* & F_0^* + F_d^* \end{pmatrix}.$$

Moreover, the blocks B, C and D satisfy $BD^d = 0$ and

$$\begin{pmatrix} BC \\ BDC \\ \vdots \\ BD^{d-1}C \end{pmatrix} = C.$$

The structure of U^* is advantageous for imposing appropriate approximation properties (sum rules [4]) on the elements of the masks p and q. The unitarity of both U and the $ABCD$-matrix lead to parametrizations of all p and q of interest in terms of few real parameters. The corresponding classes of the masks p and q include all known univariate wavelet and multi-wavelet masks.

All our results generalize easily to the case of a general dilation factor.

References

[1] J. Agler, J .E McCarthy *Pick interpolation and Hilbert function space*, volume 44 of Graduate Studies in Mathematics, AMS, Providence, RI (2002).

[2] M. Charina, M. Putinar, C. Scheiderer, J. Stöckler, *An algebraic perspective on multivariate tight wavelet frames II*, Appl. Comput. Harmon. Anal. **39** (2015), 185–213.

[3] I. Daubechies, *Ten lectures on wavelets*, in: CBMS Conf. Series in Appl. Math. 61 (1992).

[4] K. Jetter, G. Plonka, *A survey on L_2-approximation order from shift-invariant spaces*, Multivariate Approximation and Applications, N. Dyn, D. Leviatan, D. Levin, A. Pinkus (eds.), Cambridge University Press, 2001, 73–111.
[5] A. Ron, Z. Shen, *Affine systems in $L_2(\mathbb{R}^d)$: the analysis of the analysis operator*, J. Funct. Anal. **148** (1997), 408–447.
[6] J. von Neumann, *Eine Spektraltheorie für allgemeine Operatoren eines unitären Raumes*, Math. Nachr. **4** (1951), 258–281.

Density of sampling and interpolation in reproducing kernel Hilbert spaces

Karlheinz Gröchenig
(joint work with Hartmut Führ, Antti Haimi, Andreas Klotz, José Luis Romero)

How many samples of a function f are necessary to completely recover f in a given space? The first answer is the sampling theorem of Whittaker, Kotelnikov, Shannon, and others, the decisive mathematical theorems were derived by Landau [2] who gave a precise meaning of the concept of a Nyquist rate for bandlimited functions. To this day, Landau's theorem is the prototype of a density theorem, it has inspired several hundred papers on sampling. Landau's necessary conditions have been transferred, modified, and adapted to dozens of similar situations, such as sampling in spaces of analytic functions, density conditions of Gabor frames (this topic alone has attracted about hundred papers [1]), sampling in spaces of bandlimited functions on Lie groups, or the density of frames in the orbit of an irreducible unitary representation of a homogeneous nilpotent Lie group.

All these density theorems treat certain Hilbert spaces with a reproducing kernel. This fact and the similarity of all proofs raises the question of a universal density theorem in reproducing kernel Hilbert spaces. This point of view leads immediately to the following questions: What is the relevant density concept in a reproducing kernel Hilbert space? Is there a critical density in a reproducing kernel Hilbert space that separates sets of sampling from sets of interpolation?

In our contribution we attempt to give an answer and formulate a general density theorem for functions in a reproducing kernel Hilbert space. A simplified version of our main result can be stated as follows.

Theorem 1. *Let X be a metric measure space with a metric d and a measure μ. Furthermore, let $\mathcal{H} \subseteq L^2(X,\mu)$ be a reproducing kernel Hilbert space with a reproducing kernel $k_x(y) = k(y,x)$, such that $f(x) = \langle f, k_x\rangle$ for $f \in \mathcal{H}$ and $x \in X$. We impose the following* geometric conditions *on (X,d,μ):*

- *μ is non-degenerate, i.e., $\inf_x \mu(B_r(x)) > 0$ for some $r > 0$,*
- *μ is locally doubling, i.e., $\sup_{x\in X} \frac{\mu(B_{2r}(x))}{\mu(B_r(x))} < \infty$ for every $r > 0$,*
- *μ satisfies the weak annular decay property, i.e.,*

$$\lim_{r\to\infty} \sup_{x\in X} \frac{\mu(B_r(x) \setminus B_{r-1}(x))}{\mu(B_r(x))} = 0.$$

- *Compatibility of metric and measure: we have*

$$\lim_{r\to\infty}\sup_{x\in X}\int_{B_r(x)^c} d(x,y)^{-2\sigma}d\mu(y)=0$$

for some $\sigma>0$.

We impose the following off-diagonal decay condition on the reproducing kernel k *(with the same* σ*):*

$$|k(x,y)|\le C\big(1+d(x,y)\big)^{-\sigma} \qquad \text{for all } x,y\in X\,. \tag{1}$$

Then the following version of Landau's theorem holds:

(i) Necessary conditions for sampling: If for $\Lambda\subset X$ *there exist* $A,B>0$ *such that*

$$A\|f\|^2\le\sum_{\lambda\in\Lambda}|f(\lambda)|^2\le B\|f\|^2 \qquad \text{for all } f\in\mathcal{H}\,, \tag{2}$$

then

$$D^-(\Lambda):=\liminf_{r\to\infty}\inf_{x\in X}\frac{\#(\Lambda\cap B_r(x))}{\mu(B_r(x))}\ge\liminf_{r\to\infty}\inf_{x\in X}\frac{1}{\mu(B_r(x))}\int_{B_r(x)}k(y,y)d\mu(y)\,.$$

(ii) Necessary conditions for interpolation: Likewise, if for every sequence $a=(a_\lambda)_{\lambda\in\Lambda}\in\ell^2(\Lambda)$ *there exists* $f\in\mathcal{H}$*, such that* $f(\lambda)=a_\lambda,\forall\lambda\in\Lambda$*, then*

$$D^+(\Lambda):=\limsup_{r\to\infty}\sup_{x\in X}\frac{\#(\Lambda\cap B_r(x))}{\mu(B_r(x))}\le\limsup_{r\to\infty}\sup_{x\in X}\frac{1}{\mu(B_r(x))}\int_{B_r(x)}k(y,y)d\mu(y)\,.$$

Following established terminology, a set $\Lambda\subseteq X$ satisfying (2) is called a set of (stable) sampling.

The densities $D^-(\Lambda)$ and $D^+(\Lambda)$ are the obvious generalizations of the lower and upper Beurling density to metric spaces.

The principal merit of Theorem 1 is the clarification of the main notions that go into a density theorem. To prove a density theorem, one needs

(i) geometric data and the compatibility of metric and measure, and

(ii) estimates for the reproducing kernel.

The verification of these properties is by no means trivial. Indeed, kernel estimates (Bergman, Bargmann, and other reproducing kernels) constitute a deep and rich area of analysis. Theorem 1 shifts the emphasis in proofs of density theorems: it is important to understand the geometry and the reproducing kernel, but it is no longer necessary to prove a "new" density theorem from scratch with tedious modifications of known techniques.

Many known density theorems in the literature can be understood as an example of the general density theorem in reproducing kernel Hilbert spaces. To demonstrate the wide applicability of Theorem 1 we rederive some of the fundamental density theorems in several areas of analysis.

(i) *Signal analysis*: as already indicated, Theorem 1 implies Landau's necessary density conditions for bandlimited functions.

(ii) *Complex analysis* (in several variables): we will deduce Lindholm's density conditions [3] for generalized Fock spaces.

(iii) *Harmonic analysis*: We will derive a necessary condition for the density of a frame in the orbit of a square-integrable, unitary representation of a group of polynomial growth. A special case of this result is the density theorem for Gabor frames.

Theorem 1 is definitely not the end of density theorems. The weak annular decay property of the measure is not always satisfied, as it is tied to the growth of balls in X. Thus Theorem 1 excludes a number of very interesting examples, for instance, density theorems in Bergman spaces [5] and the density of wavelet frames. However, in these cases the Beurling density is not the correct density, and to this date it is an open problem whether a critical density even exists in this context.

References

[1] C. Heil. History and evolution of the density theorem for Gabor frames. *J. Fourier Anal. Appl.*, 13(2):113–166, 2007.

[2] H. J. Landau. Necessary density conditions for sampling and interpolation of certain entire functions. *Acta Math.*, 117:37–52, 1967.

[3] N. Lindholm. Sampling in weighted L^p spaces of entire functions in $\mathbb{C}^n$ and estimates of the Bergman kernel. *J. Funct. Anal.*, 182(2):390–426, 2001.

[4] K. Seip. Density theorems for sampling and interpolation in the Bargmann-Fock space. I. *J. Reine Angew. Math.*, 429:91–106, 1992.

[5] K. Seip. *Interpolation and sampling in spaces of analytic functions*, volume 33 of *University Lecture Series*. American Mathematical Society, Providence, RI, 2004.

Durrmeyer type operators with respect to arbitrary measure

Elena E. Berdysheva

Let $\mathbb{S}^d := \{x = (x_1, \dots, x_d) \in \mathbb{R}^d : 0 \le x_1, \dots, x_d \le 1,\ x_1 + \cdots + x_d \le 1\}$ be the standard simplex in $\mathbb{R}^d$. We will use the barycentric coordinates $\mathbf{x} = (x_0, x_1, \dots, x_d)$, $x_0 := 1 - x_1 - \cdots - x_d$. The d-variate *Bernstein basis polynomials* of degree n are defined by

$$B_\alpha(x) := \binom{n}{\alpha}\mathbf{x}^\alpha = \frac{n!}{\alpha_0!\alpha_1!\dots\alpha_d!}\, x_0^{\alpha_0} x_1^{\alpha_1} \cdots x_d^{\alpha_d},$$

where $\alpha = (\alpha_0, \alpha_1, \dots, \alpha_d) \in \mathbb{N}_0^{d+1}$ with $|\alpha| := \alpha_0 + \alpha_1 + \cdots + \alpha_d = n$. Clearly, B_α are non-negative on $\mathbb{S}^d$ and $\sum_{|\alpha|=n} B_\alpha(x) = 1$. Moreover, $\{B_\alpha\}_{|\alpha|=n}$ constitute a basis of the space of d-variate algebraic polynomials of total degree at most n.

The famous *Bernstein operator* is defined for $f \in C(\mathbb{S}^d)$ by

$$\big(\mathbf{B}_n f\big)(x) := \sum_{|\alpha|=n} f\left(\frac{\alpha}{n}\right) B_\alpha(x).$$

This is a positive linear operator that reproduces linear functions; $\mathbf{B}_n f$ is the sequence of polynomials of degree n that converges to f uniformly on $\mathbb{S}^d$ for

all $f \in C(\mathbb{S}^d)$. This operator, in the one-dimensional case, was introduced by S.N. Bernstein in 1912 to provide a constructive proof of the Weierstraß Approximation Theorem; this operator and its variants and modifications were studied in hundreds of papers.

One – and probably the most interesting – of the modifications of the Bernstein operator for integrable functions is the so-called *Bernstein-Durrmeyer operator* defined for $f \in L^q(\mathbb{S}^d)$, $1 \le q < \infty$, or $f \in C(\mathbb{S}^d)$ by

$$\big(\mathbf{M}_n f\big)(x) := \sum_{|\alpha|=n} \frac{\int_{\mathbb{S}^d} f(y)\, B_\alpha(y)\, dy}{\int_{\mathbb{S}^d} B_\alpha(y)\, dy}\, B_\alpha(x).$$

$\mathbf{M}_n$ is a positive linear operator that reproduces constant functions, $\mathbf{M}_n f$ convergence to f in $L^q(\mathbb{S}^d)$, $1 \le q < \infty$, or in $C(\mathbb{S}^d)$, respectively. This operator was introduced by Durrmeyer (1967) and, independently, by Lupaş (1972) and became known due to Derriennic (starting from 1981). Also the Bernstein-Durrmeyer operator with Jacobi weights was studied by a number of authors.

Here we consider the following generalization of this operator. Let ρ be a non-negative bounded (regular) Borel measure on $\mathbb{S}^d$ such that $\operatorname{supp}\rho \setminus (\partial\mathbb{S}^d) \neq \emptyset$. The *Bernstein-Durrmeyer operator with respect to the measure* ρ is defined for $f \in L^q_\rho(\mathbb{S}^d)$, $1 \le q \le \infty$, by

$$(\mathbf{M}_{n,\rho} f)(x) := \sum_{|\alpha|=n} \frac{\int_{\mathbb{S}^d} f(y)\, B_\alpha(y)\, d\rho(y)}{\int_{\mathbb{S}^d} B_\alpha(y)\, d\rho(y)}\, B_\alpha(x).$$

$\mathbf{M}_{n,\rho}$ is a positive linear operator that reproduces constant functions. It was for the first time systematically investigated in [1], to our knowledge. The motivation for this generalization came from learning theory. Our starting point was paper [2] by D.-X. Zhou and K. Jetter; they considered the univariate $\mathbf{M}_{n,\rho}$ and used it for estimates for support vector machine classifiers with polynomial kernels. Later on, B.-Z. Li [3] used the multivariate operators $\mathbf{M}_{n,\rho}$ to obtain estimates for learning rates of least-square regularized regression with polynomial kernels. Note that $\mathbf{M}_{n,\rho}$ is a compact self-adjoint integral operator in $L^2(\mathbb{S}^d, \rho)$, and its kernel is Mercer kernel. Some properties of this kernel were studied in [1].

The main topic of this talk is convergence of the operator. The results on pointwise and uniform convergence were obtained by the author in [4], [5].

Theorem. Let $x \in \operatorname{supp}\rho$. Let f be bounded on $\operatorname{supp}\rho$ and continuous at x. Then

$$\lim_{n\to\infty} |f(x) - \mathbf{M}_{n,\rho} f(x)| = 0.$$

Theorem. Let A be a compact set, $A \subset (\operatorname{supp}\rho)^\circ$. Let f be bounded on $\operatorname{supp}\rho$ and continuous on A. Then

$$\lim_{n\to\infty} \|f - \mathbf{M}^{[c]}_{n,\rho} f\|_{C(A)} = 0.$$

Convergence in the weighted L^q-spaces was proved by B.-Z. Li [3].

Theorem. Let $1 \le q < \infty$. Then

$$\lim_{n\to\infty} \|f - \mathbf{M}_{n,\rho}\, f\|_{L^q(\mathbb{S}^d,\rho)} = 0$$

for every $f \in L^q(\mathbb{S}^d, \rho)$.

She also gave estimates for the rate of convergence of $\mathbf{M}_{n,\rho}$ in the space $L^q(\mathbb{S}^d, \rho)$ in terms of certain K-functional. These estimates were improved in [6]; they play a role in studying the learning rates in the corresponding applications in learning theory.

The construction and the results can be carried over to the Szász-Mirakjan-Favard operator and the Baskakov operator [7].

Finally, in joint work in progress with K. Baumann and M. Heilmann, we consider a further generalization of the operator $\mathbf{M}_{n,\rho}$ which makes it possible to include into the same construction further operators like the Bernstein operator $\mathbf{B}_n$, the Kantorovich operator, etc., together with the Bernstein-Durrmeyer operator. Therefore, we allow the measure ρ to be different in different terms. Let $\rho = \{\rho_\alpha\}_{|\alpha|=n, n\in\mathbb{N}}$ be a collection of non-negative bounded (regular) Borel measures on $\mathbb{S}^d$ such that $\operatorname{supp}\rho_\alpha \setminus (\partial\mathbb{S}^d) \neq \emptyset$. The *Bernstein-Durrmeyer operator with respect to the collection of measures* ρ is defined by

$$(\mathbf{M}_{n,\rho}\, f)(x) := \sum_{|\alpha|=n} \frac{\int_{\mathbb{S}^d} f(y)\, B_\alpha(y)\, d\rho_\alpha(y)}{\int_{\mathbb{S}^d} B_\alpha(y)\, d\rho_\alpha(y)}\, B_\alpha(x).$$

We make our first steps in understanding assumptions on ρ and f that guarantee convergence of the operator.

References

[1] E.E. Berdysheva and K. Jetter, *Multivariate Bernstein-Durrmeyer operators with arbitrary weight functions*, J. Approx. Theory **162** (2010), 576–598.

[2] K. Jetter and D.-X. Zhou, *Approximation with polynomial kernels and SVM classifiers*, Adv. Comput. Math. **25** (2006), 323–344.

[3] B.-Z. Li, *Approximation by multivariate Bernstein-Durrmeyer operators and learning rates of least-square regularized regression with multivariate polynomial kernels*, J. Approx. Theory **173** (2013), 33–55.

[4] E.E. Berdysheva, *Uniform convergence of Bernstein-Durrmeyer operators with respect to arbitrary measure*, J. Math. Anal. Appl. **394** (2012), 324–336.

[5] E.E. Berdysheva, *Bernstein-Durrmeyer operators with respect to arbitrary measure, II: pointwise convergence*, J. Math. Anal. Appl. **418** (2014), 734–752.

[6] E.E. Berdysheva and B.-Z. Li, *On L^p-convergence of Bernstein-Durrmeyer operators with respect to arbitrary measure*, Publ. Inst. Math., Nouv. Sér. **96(110)** (2014), 23–29.

[7] E.E. Berdysheva and E. Al-Aidarous, *Szász-Mirakjan-Durrmeyer and Baskakov-Durrmeyer operators with respect to arbitrary measure*, Jaen J. Approx., to appear.

Reporter: Andreas Christmann

Participants

Prof. Dr. Elena Berdysheva
Mathematisches Institut
Justus-Liebig-Universität Gießen
Arndtstrasse 2
35392 Gießen
GERMANY

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

Prof. Dr. Stephane Boucheron
UFR de Mathématiques
Université Paris Diderot - Paris VII
5, rue Thomas Mann
75205 Paris Cedex 13
FRANCE

Prof. Dr. Martin D. Buhmann
Lehrstuhl für Numerische Mathematik
Universität Gießen
Arndtstrasse 2
35392 Giessen
GERMANY

Dr. Maria Charina
Institut für Mathematik
Universität Wien
1090 Wien
AUSTRIA

Prof. Dr. Andreas Christmann
Fakultät für Mathematik, Physik und Informatik
Lehrstuhl für Stochastik
Universität Bayreuth
95440 Bayreuth
GERMANY

Prof. Dr. Karlheinz Gröchenig
Fakultät für Mathematik
Universität Wien
Oskar-Morgenstern-Platz 1
1090 Wien
AUSTRIA

Prof. Dr. Xin Guo
Department of Applied Mathematics
The Hong Kong Polytechnic University
TU 825, Yip Kit Chuen Building
Hung Hom
Hong Kong
CHINA

Prof. Dr. Kurt Jetter
Institut für Angewandte Mathematik u. Statistik
Universität Hohenheim
70593 Stuttgart
GERMANY

Prof. Dr. Philipp Kügler
Institut für Angewandte Mathematik u. Statistik
Universität Hohenheim
70593 Stuttgart
GERMANY

Prof. Dr. Sayan Mukherjee
Department of Statistical Sciences
Institute for Genome Sciences & Policy
Duke University
112 Old Chemistry Bldg., Box 90251
Durham NC 27710
UNITED STATES

Prof. Dr. Gerlind Plonka-Hoch
Institut f. Numerische & Angew. Mathematik
Universität Göttingen
Lotzestrasse 16-18
37083 Göttingen
GERMANY

Prof. Dr. Tomaso Poggio
Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
77 Massachusetts Avenue
Cambridge, MA 02139
UNITED STATES

Prof. Dr. Holger Rauhut
Lehrstuhl für Mathematik C (Analysis)
RWTH Aachen
Pontdriesch 10
52062 Aachen
GERMANY

Prof. Dr. Tomas Sauer
Fakultät f. Mathematik u. Informatik
Universität Passau
94030 Passau
GERMANY

Prof. Dr. Bernhard Schölkopf
Max-Planck-Institut f. Intelligente Systeme
Office 211
Spemannstraße 38
72076 Tübingen
GERMANY

Prof. Dr. Steve Smale
Department of Mathematics
City University of Hong Kong
83 Tat Chee Avenue, Kowloon
Hong Kong
CHINA

Prof. Dr. Gabriele Steidl
Fachbereich Mathematik
Technische Universität Kaiserslautern
67653 Kaiserslautern
GERMANY

Prof. Dr. Ingo Steinwart
Fachbereich Mathematik
Universität Stuttgart
Pfaffenwaldring 57
70569 Stuttgart
GERMANY

Prof. Dr. Joachim Stöckler
Institut für Angewandte Mathematik
Technische Universität Dortmund
44221 Dortmund
GERMANY

Prof. Dr. Johan Suykens
Department of Electrical Engineering
KU Leuven ESAT/STADIUS
Office B00.16, bus 2446
Kasteelpark Arenberg 10
3001 Leuven
BELGIUM

Prof. Dr. Alexandre B. Tsybakov
CREST
Timbre J 340
3, Avenue P. Larousse
92240 Malakoff Cedex
FRANCE

Prof. Dr. Holger Wendland
Mathematisches Institut
Universität Bayreuth
95440 Bayreuth
GERMANY

Dr. Daohong Xiang
Mathematisches Institut
Universität Bayreuth
95440 Bayreuth
GERMANY

Prof. Dr. Yiming Ying
Department of Mathematics
State University of New York at Albany
1400 Washington Avenue
Albany, NY 12222
UNITED STATES

Prof. Dr. Ding-Xuan Zhou
Department of Mathematics
City University of Hong Kong
83 Tat Chee Avenue, Kowloon
Hong Kong
CHINA