Mathematisches Forschungsinstitut Oberwolfach

Report No. 31/2012

DOI: 10.4171/OWR/2012/31

Learning Theory and Approximation

Organised by
Kurt Jetter, Hohenheim
Steve Smale, Hong Kong
Ding-Xuan Zhou, Hong Kong

June 24th – June 30th, 2012

Abstract. Learning theory studies data structures from samples and aims at understanding unknown function relations behind them. This leads to interesting theoretical problems which can be often attacked with methods from Approximation Theory. This workshop - the second one of this type at the MFO - has concentrated on the following recent topics: Learning of manifolds and the geometry of data; sparsity and dimension reduction; error analysis and algorithmic aspects, including kernel based methods for regression and classification; application of multiscale aspects and of refinement algorithms to learning.

Mathematics Subject Classification (2000): 68Q32, 41A35, 41A63, 62Jxx.

Introduction by the Organisers

The workshop *Learning Theory and Approximation*, organised by Kurt Jetter (Stuttgart-Hohenheim), Steve Smale (Hong Kong) and Ding-Xuan Zhou (Hong Kong) was held June 24–30, 2012. The meeting was well attended with 47 participants from Asia, Europe and North America. It provided an excellent platform for fruitful interactions among scientists from learning theory and approximation theory.

The first part of the scientific program consisted of a few talks on learning geometric structures from data. Steve Smale's talk on mathematical foundations of immunology demonstrated strong connections on data analysis for peptides and amino acid chains among the research fields of computational biology and geometry, learning theory and approximation theory. Nat Smale presented a Hodge

theory for Alexandrov spaces with curvature bounded above including Riemannian manifolds, Riemannian manifolds with boundary and singular spaces such as simplicial complexes whose faces have constant curvature, and Titz buildings. The described Hodge decomposition can be applied to data analysis and processing. Modelling data by manifolds reflects many important aspects of realistic data and provides a direct connection with differential geometry. In particular, manifold learning algorithms based on graph Laplacians constructed from data have received considerable attention both in practical applications and theoretical analysis. Belkin discussed the behavior of graph Laplacians at points at or near boundaries, intersections and edges, and showed that the behavior of graph Laplacians near these singularities is quite different from that in the interior of the manifolds. Jost spoke about the topic of geometric structures on the space of probability measures in the area of information geometry, and introduced a sufficient statistic for a parametrized family of measures under which the Fisher metric and the Amari-Chentsov tensors remain invariant. Von Luxburg talked about the problem of density estimation from unweighted k-nearest neighbor graphs, and connections to graph learning algorithms like spectral clustering or semi-supervised learning. Lim gave a talk on principal components of cumulants, discussing the geometry underlying cumulants and examining two ways to their principal components analysis, decomposing a homogeneous form into a linear combination of powers of linear forms, and decomposing a symmetric tensor into a multilinear combination of points on a Stiefel manifold.

Sparsity is an important property for dimension reduction, data representation and analysis, and information retrieval. In this workshop, some statisticians and approximation theorists discussed sparsity for various purposes and raised interesting problems for approximation theory: Tsybakov introduced a compound functional model as a nonparametric generalization of the high-dimensional linear regression model under the sparsity scenario and presented minimax rates of convergence in terms of structural conditions of functions. Dahmen applied deep analysis from tree-structured approximation to classification algorithms with adaptive partitioning and analyzed their risk performance. The analysis allows one to relax classical Hölder smoothness to weaker Besov smoothness, which leads to interesting approximation theory problems. Sparsity was a core issue in classical support vector machines. Christmann's talk was focussed on the question how to draw statistical decisions based on nonparametric methods such as bootstrap approximations of support vector machines and qualitatively robust support vector machines. His discussions on various loss functions raised research problems for approximation theorists. Li considered the compressed sensing topic of nonuniform support recovery via orthogonal matching pursuit from noisy random measurements. Zhou talked about error analysis and sparsity for support vector regression, coefficient-based regularization with ℓ^1-penalty, and kernel projection machines with ℓ^q-penalty.

Both approximation theory and learning theory provide useful tools for data analysis and statistics. This is reflected by quite a few talks in this workshop.

Wu gave a learning theory perspective on the empirical minimum error entropy (MEE) principle developed in the fields of signal processing and data mining, and he provided a rigorous consistency analysis of some MEE learning algorithms in terms of approximation theory conditions on the model and hypothesis spaces. Döring considered a regression model with a change point in the regression function and investigated the consistency with the increasing sample size of the least squares estimates of the change point, for which the convergence rates depend on the order of smoothness of the regression function at the change point. Minh proposed a regularized spectral algorithm for hidden Markov models with numerical stability, and gave some theoretical justification and simulations on real data from pattern recognition. Pereverzyev applied the least squares Tikhonov regularization schemes in reproducing kernel Hilbert spaces to the practical problem of blood glucose reading, discussed intensively how to choose the hypothesis space, and described a kernel adaptive regularized algorithm.

Kernels have been an essential part of both learning theory and approximation theory. They form the topic of a few talks in this workshop. Plonka introduced Prony's method for solving inverse problems to the workshop audience, and described her recent work on function reconstruction in terms of sparse Legendre expansions and a new perception of Prony's method based on eigenfunctions of linear operators. Steinwart surveyed approximation theory properties of reproducing kernel Hilbert spaces (eigenvalues, entropy numbers, interpolation spaces, Mercer representations) and some related kernel methods for both supervised and unsupervised learning. Schaback described some methods for explicit constructions of new positive definite radial kernels, in particular kernels that are linked to generalized Sobolev spaces. Zu Castell demonstrated some kernel-based methods for learning and approximation. For conditionally positive definite kernels, he raised some approximation theory questions about the associated reproducing kernel Pontryagin space. Rosasco described the problem of learning the region where a probability measure is concentrated by means of separating kernels. Some approximation theory questions are mentioned such as the approximation of sets under the Hausdorff distance and the existence of completely separating kernels.

Approximation theory and ideas of multiscale analysis from wavelets have been applied in learning theory and have further potential applications. The workshop contains quite a few talks discussing these areas and other possible connections between learning and approximation. Kutyniok gave a survey on shearlets and demonstrated their applications in sparse approximation and dictionary learning. Mhaskar talked about function approximation on data dependent manifolds. The research area of irregular sampling was described by Stöckler in his talk. Han's talk was on linear-phase moments in wavelet analysis and approximation theory. Bernstein polynomials and Bernstein-Durrmeyer operators associated with general probability measures together with their applications to learning theory were discussed by Wu and Berdysheva in their talks. The ideas of tracking multiscale structures by subdivision schemes and refinement algorithms together with potential applications in learning theory were discussed by Ebner and Jetter.

The organizers acknowledge the friendly atmosphere provided by the Oberwolfach institute, and express their thanks to the entire staff.

Workshop: Learning Theory and Approximation

Table of Contents

Abstracts

Mathematical foundation of immunology

Steve Smale

Large scientific and industrial enterprises are engaged in efforts to produce new vaccines from synthetic peptides. The study of peptide binding to appropriate alleles is a major part of this effort [2]. Our goal here [1] is to support the use of a certain string kernel for peptide binding prediction as well as for the classification of supertypes of the major histocompatibility complex (MHC, in humans which is also called HLA) alleles.

Peptide binding to a fixed HLAII (and HLAI as well) molecule (or an allele) a is a crucial step in the immune response of the human body to a pathogen or a peptide-based vaccine. Its prediction is computed from data of the form $\{(x_i, y_i)\}_{i=1}^m$ with $x_i \in \mathcal{P}_a$ and $y_i \in [0,1]$ where $\mathcal{P}_a$ is a set of peptides (i.e., chains of amino acids or peptides of length 9 to 37, usually about 15) associated to an HLAII allele a. The peptide binding problem occupies much research. To study this problem we may use a symmetric function (a kernel) $K : X \times X \to \mathbb{R}$ where X is a finite set ($\mathcal{P}_a \subseteq X$ for the peptide binding problem with a single allele a). Given an order on X, K may be represented as a matrix. Then it is assumed that K is positive definite, and it generates a reproducing kernel Hilbert space $\mathcal{H}_K$.

Following regularized least squares (RLS) supervised learning [3], the main construction is to compute

$$f_a = \arg\min_{\mathcal{H}_K} \left\{ \sum_{i=1}^{m} (f(x_i) - y_i)^2 + \lambda \|f\|_K^2 \right\}. \tag{1}$$

Here $\lambda > 0$ is a regularization parameter chosen by a procedure called leave-one-out cross validation.

There is an important generalization of the peptide binding problem where the allele is allowed to vary [1].

The construction of our main kernel K on amino acid chains, denoted as $\widehat{K}^3$ later, plays an essential role in our study. It is inspired by local alignment kernels as well as an analogous kernel in vision.

Let $\mathcal{A}$ be the set of the 20 basic (for life) amino acids. Every protein has a representation as a string of elements of $\mathcal{A}$. Our construction of the kernel is given in three steps [1].

Step 1. Definition of a kernel $K^1 : \mathcal{A} \times \mathcal{A} \to \mathbb{R}$. BLOSUM62 is a similarity (or substitution) matrix on $\mathcal{A}$ frequently used in immunology [4]. In the formulation of BLOSUM62, a kernel $Q : \mathcal{A} \times \mathcal{A} \to \mathbb{R}$ is defined using blocks of aligned strings of amino acids representing proteins. We define a BLOSUM62-2 matrix, indexed by the set $\mathcal{A}$, by normalizing Q and taking a power $\beta > 0$.

Step 2. Let $\mathcal{A}^1 = \mathcal{A}$ and define $\mathcal{A}^{k+1} = \mathcal{A}^k \times \mathcal{A}$ recursively for any $k \in \mathbb{N}$. We say s is an amino acid chain (or string) if $s \in \cup_{k=1}^{\infty} \mathcal{A}^k$, and $s = (s_1, \ldots, s_k)$ is a k-mer if $s \in \mathcal{A}^k$ for some $k \in \mathbb{N}$ with $s_i \in \mathcal{A}$. Consider

$$K_k^2(u, v) = \Pi_{i=1}^k K^1(u_i, v_i)$$

where u, v are amino acid strings of the same length k, $u = (u_1, \ldots, u_k), v = (v_1, \ldots, v_k)$; u, v are k-mers. K_k^2 is a kernel on the set of all k-mers.

Step 3. Let $f = (f_1, \ldots, f_m)$ be an amino acid chain. Denote $|f|$ as the length of f (so here $|f| = m$). Write $u \subset f$ whenever u is of the form $u = (f_{i+1}, \ldots, f_{i+k})$ for some $1 \leq i + 1 \leq i + k \leq m$. Let g be another amino acid chain, then define

$$K^3(f, g) = \sum_{\substack{u \subset f, v \subset g \\ |u| = |v| = k \\ \text{all } k = 1, 2, \ldots}} K_k^2(u, v).$$

We define the desired correlation kernel $\widehat{K}^3$ by normalizing the above kernel

$$\widehat{K}^3(x, y) = \frac{K^3(x, y)}{\sqrt{K^3(x, x) K^3(y, y)}}.$$

References

[1] Wen-Jun Shen, Hau-San Wong, Quan-Wu Xiao, Xin Guo, and Stephen Smale, *Towards a mathematical foundation of Immunology and Amino Acid Chains*, preprint (2012).

[2] O. Lund, M. Nielsen, C. Lundegaard, C. Kesmir, and S. Brunak, *Immunological Bioinformatics*, The MIT Press (2005).

[3] S. Smale and D. X. Zhou, *Learning theory estimates via integral operators and their approximations*, Constr. Approx. **26** (2007), 153–172.

[4] S. Henikoff and J. G. Henikoff, *Amino acid substitution matrices from protein blocks*, Proceedings of the National Academy of Sciences **89** (1992), 10915–10919.

A Hodge theory for Alexandrov spaces with curvature bounded above

Nat Smale

In previous joint work with Laurent Bartholdi, Thomas Schick and Steve Smale in [1], a Hodge theory for compact metric spaces was proposed and partially developed. A fundamental aspect of this theory, is that it describes a cohomology at a fixed scale. Let (X, d) be a compact metric space, which we assume is endowed with a Borel probability measure μ, and let $\alpha > 0$ (the scale). For $k = 1, 2, \ldots$, we define $U_\alpha^{k+1} \subset X^{k+1}$ to be the closed α-neighborhood of the diagonal in X^{k+1}. In our theory, the set of cochains of degree k, analogous to the differential k-forms in the classical theory on a smooth manifold, is the Hilbert space of alternating functions on U_α^{k+1} which are in L^2, denoted by $L_a^2(U_\alpha^{k+1})$. The differential

$\delta : L_a^2(U_\alpha^{k+1}) \to L_a^2(U_\alpha^{k+2})$ is the Alexander-Spanier coboundary operator

$$\delta f(x_0, \ldots, x_{k+1}) = \sum_{i=0}^{k+1} (-1)^{i+1} f(x_0, \ldots, \hat{x}_i, \ldots, x_{k+1}) \; .$$

It is easily seen that δ defines a bounded linear map, and that $\delta^2 = 0$, and therefore gives rise to a cochain complex of Hilbert spaces

$$0 \longrightarrow L^2(X) \xrightarrow{\delta_0} L_a^2(U_\alpha^2) \xrightarrow{\delta_1} \cdots \xrightarrow{\delta_{k-1}} L_a^2(U_\alpha^{k+1}) \xrightarrow{\delta_k} \cdots$$

We view this as analogous to the de Rham complex, and the corresponding cohomology $H^k_{\alpha,L^2} = \frac{Ker\,\delta}{Im\,\delta}$ describes a cohomology at scale α. The adjoint $\delta^* : L_a^2(U_\alpha^{k+2}) \to L_a^2(U_\alpha^{k+1})$ is given by

$$\delta^* f(x_0, \ldots, x_k) = (k+2) \int_{S_x} f(t, x) \, d\mu(t)$$

where $S_x = \{t \in X : (t, x) \in U_\alpha^{k+2}\}$ is the slice of $x = (x_0, \ldots, x_k)$. The corresponding Hodge Laplacian $\Delta : L_a^2(U_\alpha^{k+1}) \to L_a^2(U_\alpha^{k+1})$ is given by

$$\Delta = \delta\delta^* + \delta^*\delta \; .$$

In [1], various conditions were given on X, d, μ, α which imply that the Hodge decomposition holds:

$$L_a^2(U_\alpha^{k+1}) = Im\,\delta \oplus Im\,\delta^* \oplus Ker\,\Delta$$
$$\text{and } Ker\,\Delta \text{ is isomorphic to } H^k_{\alpha,L^2} = \tfrac{Ker\,\delta}{Im\,\delta}$$

To be more precise, let $\mathcal{K}(X)$ denote the metric space of nonempty compact subsets of X endowed with the Hausdorff metric, and define the witness function

$$w_\alpha : U_\alpha^{k+1} \to \mathcal{K}(X)$$

by $w_\alpha(x_0, \ldots, x_k) = \cap_i B_\alpha(x_i)$ where $B_\alpha(x_i)$ is the closed ball of radius α centered at x_i. It was shown that if w_α was continuous, and that the radius of finite intersections of balls of radius $\alpha + \delta$ (δ sufficiently small) is less than $\alpha + \delta$, then the corresponding Hodge decomposition holds. In particular, it was shown that if X was a compact Riemannian manifold, and α was sufficiently small, these conditions hold.

Here, we extend these results to a large class of metric spaces, namely Alexandrov spaces with curvature bounded above. These are geodesic spaces, where distances between nearby points are realized by the length of a path (a geodesic), and whose geodesic triangles satisfy a certain comparison with triangles in a space form of constant curvature K (for some fixed $K \in \mathbf{R}$). Examples of Alexandrov spaces with curvature bounded above include Riemannian manifolds, Riemannian manifolds with boundary, as well as singular spaces such as simplicial complexes whose faces have constant curvature, and Titz buildings. It is shown that if X is a compact Alexandrov space with curvature bounded above, and $\alpha > 0$ is sufficiently small, then the sufficient conditions described above hold, and thus the Hodge decomposition follows.

References

[1] L. Bartholdi, T. Schick, N. Smale and S. Smale, *An abstract Hodge theory*, submitted to J.F.O.C.M.

Dealing with singular manifolds: theory and applications

Mikhail Belkin

(joint work with Qichao Que, Yusu Wang, Xueyuan Zhou)

An important setting for many recent algorithms and analyses in machine learning problem has been that the underlying data lies on or near a smooth embedded manifold. This model reflects many important aspects of realistic data and provides a direct connection with classical differential geometry.

At the same time, it can be argued that singularities and boundaries are an important aspect of the geometry of realistic data. Boundaries occur whenever the process generating data has a bounding constraint; while singularities appear when two different manifolds intersect or if a process undergoes a "phase transition", changing non-smoothly as a function of a parameter.

In manifold learning, algorithms based on graph Laplacian constructed from data have received considerable attention both in practical applications and theoretical analysis. Much of the existing work has been done under the assumption that the data is sampled from a manifold without boundaries and singularities or that the functions of interest are evaluated away from such points.

In this talk I will discuss the behavior of graph Laplacians at points at or near boundaries and two main types of other singularities: *intersections*, where different manifolds come together and sharp *"edges"*, where a manifold sharply changes direction. We show that the behavior of graph Laplacian near these singularities is quite different from that in the interior of the manifolds. In fact, a phenomenon somewhat reminiscent of the Gibbs effect in the analysis of Fourier series, can be observed in the behavior of graph Laplacian near such points. Unlike in the interior of the domain, where graph Laplacian converges to the Laplace-Beltrami operator, near singularities graph Laplacian tends to a first-order differential operator, which exhibits different scaling behavior as a function of the kernel width. One important implication is that while points near the singularities occupy only a small part of the total volume, the difference in scaling results in a disproportionately large contribution to the total behavior. Another significant finding is that while the scaling behavior of the operator is the same near different types of singularities, they are very distinct at a more refined level of analysis.

I will argue that a comprehensive understanding of these structures in addition to the standard case of a smooth manifold can take us a long way toward better methods for analysis of complex non-linear data and can lead to significant progress in algorithm design.

In particular, I will describe a recent application of these methods to the problem of reconstructing a surface with sharp corners in graphics.

Information geometry and sufficient statistics

Jürgen Jost
(joint work with Nihat Ay, Hông Vân Lê and Lorenz Schwachhöfer)

Information geometry provides a geometric structure on the space of probability measures on a given space Ω. This space of probability measures can be seen as a projective space for the space of all nonnegative measures, and it thereby inherits a metric from the latter. This is the Fisher metric. Moreover, when μ_0 is some base measure, and $\phi \in L^1(\Omega, \mu_0)$, then $\mu = \phi\mu_0$ is another measure whose L^1-functions f are of the form $g + \log \phi$ for an $L^1(\Omega, \mu_0)$-function g. Since this shift by $\log \phi$ does not depend on f or g, we obtain an affine structure on the space of measures when we consider $L^1(\Omega, \mu)$ as the tangent space at μ. This affine structure was independently discovered and explored by Amari and Chentsov. There is the problem, however, that when $f \in L^1(\Omega, \mu)$, then this does not imply that also $e^f \in L^1(\Omega, \mu)$ (the exponential is the inverse of the logarithm that appeared in the transition from the function ϕ characterizing the transition between measures and the function $\log \phi$ that expressed the affine transformation between the tangent spaces involved). Thus, certain infinitesimal deformations are obstructed.

A sufficient statistic for a parametrized family of measures on Ω is given by a map into another measurable space Ω' that does not loose any information about the family parameter x. Therefore, also the Fisher metric and the Amari-Chentsov tensors remain invariant under sufficient statistics. Conversely, Chentsov showed that, for a *finite* space Ω, these tensors are uniquely characterized (up to a constant, of course) by invariance under sufficient statistics. The extension to general spaces Ω turned out to be difficult because of the above topological problems, and this had remained an open problem. In our recent work, however, we found a functorial approach to these topological aspects and could show the uniqueness of the Fisher and Amari-Chentsov tensors for invariance under sufficient statistics for any Ω.

A Learning Theory perspective on the empirical minimum error entropy principle

Qiang Wu
(joint work with Jun Fan, Ting Hu, and Ding-Xuan Zhou)

Information theoretical learning is an important research area in machine learning. It uses the concepts of entropies from information theory to substitute the conventional statistical descriptors of variances and covariances. Among various algorithms falling into this framework, the minimum error entropy (MEE) algorithm was introduced for supervised learning and applicable to both regression and classification problems. Although MEE have been successful in many applications its mathematical foundation is far from well understood. Our purpose is a rigorous consistency analysis of MEE algorithm and interpret its empirical effectiveness from a learning theory perspective.

We focus on MEE algorithm for regression problem where the aim is to estimate a target function f^* for which a set of observations $\mathbf{z} = \{(x_i, y_i) : \ i = 1, \ldots, n\}$ are obtained from a model

$$Y = f^*(X) + \epsilon, \qquad \mathbf{E}(\epsilon|X) = 0.$$

MEE algorithm associated to the Rényi's entropy of order 2 is motivated by minimizing Rényi's entropy of the error variable $E = Y - f(X)$. Let $p_{_E}$ denote the probability density function of E. The Rényi's entropy of the error variable E is defined by

$$\mathcal{R}(f) = -\log\left(\mathbf{E}[p_{_E}]\right) = -\log\left(\int_{\mathbb{R}} (p_{_E}(e))^2\, de\right).$$

In practice $p_{_E}$ can be estimated from the samples $e_i = y_i - f(x_i)$ by a kernel density estimator. Using the Gaussian kernel $G_h(t) = \frac{1}{\sqrt{2\pi}h}\mathbf{e}^{-\frac{t^2}{2h^2}}$ with bandwidth parameter h, the empirical estimation of $p_{_E}$ is given as

$$p_{_{E,\mathbf{z}}}(e) = \frac{1}{n}\sum_{j=1}^{n} G_h(e - e_j) = \frac{1}{n}\sum_{j=1}^{n} \frac{1}{\sqrt{2\pi}h}\mathbf{e}^{-\frac{(e-e_j)^2}{2h^2}}.$$

The empirical MEE algorithm learns $f_{\mathbf{z}}$ from a set of hypothesis space $\mathcal{H}$ by minimizing the empirical version of the Rényi's entropy

$$\mathcal{R}_{\mathbf{z}}(f) = -\log\left(\frac{1}{n}\sum_{i=1}^{n} p_{_{E,\mathbf{z}}}(e_i)\right) = -\log\left(\frac{1}{n^2}\sum_{i,j=1}^{n} G_h(e_i - e_j)\right).$$

That is, $f_{\mathbf{z}} = \arg\min_{f\in\mathcal{H}} \mathcal{R}_{\mathbf{z}}(f)$.

In order to study the asymptotical properties of the MEE algorithm we define two types of consistency. The error entropy consistency refers to the convergence of $\mathcal{R}(f_{\mathbf{z}})$ to $\mathcal{R}^* = \inf \mathcal{R}(f)$ in probability as $n \to \infty$. The regression consistency refers to the convergence of $f_{\mathbf{z}}$ plus a suitable constant adjustment to the regression function f^* in probability. This definition is natural because the solution of the empirical MEE algorithm is invariant to constant adjustment. Note that the error entropy consistency ensures the learnability of minimum error entropy, as is expected from the motivation of the empirical MEE algorithm, while the regression function consistency enables good approximation of the regression target function f^*. These two types of consistency, however, are not necessarily coincident.

Our main contributions are to show the incoincidence of these two types of consistency and illustrate the complication of the regression consistency. A couple of main results are proved under mild conditions on the model and hypothesis space. Firstly we show that the error entropy consistency is always true by choosing the bandwidth parameter h to tend to 0 at an appropriate slow rate. This is somewhat an expected result from the motivation of MEE algorithm. Next we studied the relation between error entropy consistency and regression consistency. For homoskedastic models where the noise ϵ is independent of X, it is proved that the error entropy consistency implies the regression consistency. As

a corollary, the regression consistency always holds if h tends to 0 at an appropriate slow rate. For heteroskedastic models where the noise variable depends on X, we presented a counter-example for which the error entropy consistency and regression function consistency do not coincide. Lastly, we prove a quite surprising result which states that regression consistency is true for both homoskedastic and heteroskedastic models if the bandwidth parameter tends to infinity at a slow rate. The requirement of choosing large h was observed in some earlier empirical work but clearly contradicts the motivation of MEE algorithms because the kernel density estimator does not converge without $h \to 0$. These results show that the consistency of the empirical MEE is a very complicated issue and requires further investigation.

Change point estimation in regression models

Maik Döring
(joint work with Uwe Jensen)

In this talk we consider a simple regression model with a change point in the regression function. We investigate the consistency with increasing sample size n of the least square estimates of the change point. It turns out that the rates of convergence depend on the order of smoothness of the regression function at the change point. In the case of a discontinuity point of a regression function we have that the rate of convergence is n. In addition, it is shown, that for the discontinuity case the change point estimator converges to a maximizer of a random walk. In the case of a smooth change of a regression function the change point estimator converges to a maximizer of a Gaussian process. The asymptotic normality property of the change point estimator is established in a particular case of a smooth change point. What goes beyond the results published in the literature so far is in particular that the rate of convergence is even valid for low degrees of smoothness of the change point.

The problem to estimate the location of a change point in a regression model has been studied in the literature to some extent. In most cases locating a jump discontinuity is considered and properties of the estimators are studied. Müller [4] investigates the problem of estimating a jump change point in the derivative of some order $\nu \geq 0$ of the regression function. His change point estimators are based on one-sided kernels. This includes the case of continuous regression functions with a change in the derivative at same point which we call smooth change point.

Let for $n \in \mathbb{N}$ the observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ be i.i.d. $\mathbb{R}^2$-valued random variables with the same distribution as (X, Y). We assume that the response variables Y_i are given by the following regression model with change point $\theta_0 \in [0, 1]$

$$Y_i = f_{\theta_0}(X_i) + \epsilon_i, \quad 1 \leq i \leq n,\ n \in \mathbb{N}.$$

For $\theta \in [0,1]$ the regression function $f_\theta : [0,1] \to \mathbb{R}$ is given by

$$f_\theta(x) := (x-\theta)^q \cdot 1_{[\theta,1]}(x),$$

where $q \geq 0$ and 1_A is the indicator function of a set A. Let $\epsilon_1, \cdots, \epsilon_n$ be i.i.d. real valued random variables with $E(\epsilon_1|X) = 0$ a.s. and suitably integrable. The case $q = 0$, i.e. the regression function has a jump at θ, was studied in Kosorok [3]. The case $q \geq 2$ was considered by Rukhin and Vajda [5] in a fixed design model. In a similar model an estimator for a singularity of a density function was analyzed in the book of Ibragimov and Has'minskii [2].

In the following the focus will be on estimating the change point θ_0 by the least squares method. We assume that the regression function is known except the change point $\theta_0 \in [0,1]$. We consider the least squares error for any possible change point. For $\theta \in [0,1]$ and $n \in \mathbb{N}$ we define

$$M_n(\theta) := -\frac{1}{n}\sum_{i=1}^{n} (Y_i - f_\theta(X_i))^2.$$

For $n \in \mathbb{N}$ our estimator is defined as the maximizing point of M_n:

$$\hat{\theta}_n := \operatorname*{argmax}_{\theta\in[0,1]} M_n(\theta).$$

To analyze the asymptotic behavior of our estimator, we use the theory of M-estimators and empirical processes, which are described, for example, in van der Vaart [6]. We show that our estimator is strongly consistent. It turns out that the rates of convergence depend on the order of smoothness q of the regression function at the change point.

$$r_n\left(\hat{\theta}_n - \theta_0\right) = O_p(1), \qquad \text{where } r_n = \begin{cases} n^{\frac{1}{2q+1}} & 0 \leq q < \frac{1}{2} \\ \sqrt{n \cdot \ln(n)} & q = \frac{1}{2} \\ \sqrt{n} & \frac{1}{2} < q < \infty. \end{cases}$$

In addition, it is shown, that for the discontinuity case the change point estimator converges to a maximizer of a random walk. In the case of a smooth change of a regression function the change point estimator converges to a maximizer of a Gaussian process. The asymptotic normality property of the change point estimator is established for $q \geq \frac{1}{2}$. What goes beyond the results published in the literature so far is in particular that the rate of convergence is even valid for low degrees of smoothness of the change point.

References

[1] M. Döring and U. Jensen, *Change point estimation in regression models with fixed design*, in: N. Rykov, N. Balakrishnan, M. S. Nikulin, *Mathematical and Statistical Models and Methods in Reliability: Applications to Medicine, Finance, and Quality Control*, pp. 207–221, Birkhäuser, Boston (2010).

[2] I. A. Ibragimov and R.Z. Has'minskii, *Statistical Estimation – Asymptotic Theory*, Springer, New York (1981).

[3] M.R. Kosorok, *Introduction to Empirical Processes and Semiparametric Inference*, Springer, New York (2008).

[4] H. G. Müller, *Change-point in nonparametric regression analysis*, The Annals of Statistics **20** (1992), 737–761.
[5] A. L. Rukhin and I. Vajda, *Change-point estimation as a nonlinear regression problem*, Statistics **30** (1997), 181–200.
[6] A. W. Van der Vaart, *Asymptotic Statistics*, Cambridge Series in Statistical and Probabilistic Mathematics (1998).

Shearlets: sparse approximation and dictionary learning

Gitta Kutyniok
(joint work with Jakob Lemvig, Wang-Q Lim)

Many important problem classes in applied mathematics are governed by anisotropic features such as singularities on lower dimensional embedded manifolds. Examples are shock fronts in solutions of transport dominated equations or edges in images. Shearlet analysis might by now be considered the most versatile and successful methodology to efficiently represent such features, in particular, because it allows a unified treatment of the continuum and digital realm. For a survey on shearlets we refer to [2]. However, although compact support is often required to achieve superior spatial localization, most research had so far focussed on band-limited shearlets.

About two years ago, an extensive theory of compactly supported shearlets was introduced in [1]. In [3], those shearlets could in fact also be shown to provide optimally sparse approximations of anisotropic features within the model situation of what are typically coined cartoon-like functions, i.e., coarsely speaking functions supported on the unit square which are C^2 apart from a closed C^2 discontinuity curve. Very recently, this theory was extended in [4] in two ways: First, it was generalized to the 3D setting, which is the first setting in which anisotropic features appear in two different dimensions. Second, the model class was extended by allowing the function as well as the curve to (independently) have a regularity of C^α with $1 < \alpha \leq 2$. The second extension required a generalization coined *hybrid shearlets* of classical shearlet systems. Those new systems can be regarded as a parameterized family of systems which range from shearlets, i.e., parabolically scaled systems, to wavelets, i.e., isotropically scaled systems.

An essential problem when utilizing systems for sparse recovery is the design of such. For this task, systems can be categorized in two classes: One class are specifically designed systems such as wavelets and shearlets, whereas the other class are data adapted systems which are *learned* from given test data by dictionary learning algorithms. The problem with systems belonging to the second class lies in the missing structure, which typically prevents a rigorous mathematical analysis of their properties. Hybrid shearlets are one possible way to bridge this gap by providing a family of functions dependent on one parameter, which can be learned. The difference to customarily exploited dictionary learning algorithms lies in the fact that the learned system is then still highly structured and, for instance, frame properties as well as sparse approximation results are known.

References

[1] P. Kittipoom, G. Kutyniok, and W.-Q. Lim, *Construction of compactly supported shearlet frames*, Constr. Approx. **35** (2012), 21–72.

[2] G. Kutyniok and D. Labate, eds., *Shearlets: Multiscale Analysis for Multivariate Data*, Birkhäuser, Boston (2012).

[3] G. Kutyniok and W.-Q. Lim, *Compactly supported shearlets are optimally sparse*, J. Approx. Theory **163** (2011), 1564–1589.

[4] G. Kutyniok, J. Lemvig, and W.-Q. Lim, *Optimally sparse approximations of 3D functions by compactly supported shearlet frames*, SIAM J. Math. Anal., to appear.

Approximation by k-sparse sums of eigenfunctions of linear operators

Gerlind Plonka

(joint work with Thomas Peter)

In signal analysis, there often is some a priori knowledge about the underlying structure of the wanted signal. Thus, one is faced with the problem of extracting a certain number of parameters from the given signal measurements. Considering for example a structured function of the form

$$f(\omega) = \sum_{j=1}^{k} c_j \, e^{\omega T_j}$$

with complex parameters c_j and T_j, $j = 1, \ldots, k$, and assuming that $-\pi < \mathrm{Im} T_1 < \ldots < \mathrm{Im} T_k < \pi$, one aims to reconstruct c_j and T_j from a given small amount of (possibly noisy) measurement values $f(\ell)$. Using Prony's method or one of its stabilized variants, one is able to reconstruct f with only $2k$ function values $f(\ell)$, $\ell = 0, \ldots, 2k-1$. The solution of this problem involves the determination of a so-called "Prony polynomial"

$$\Lambda(z) = \prod_{j=1}^{k} (z - e^{T_j}) = \sum_{\ell=0}^{k} \alpha_\ell z^\ell$$

with $\alpha_k = 1$. Using the structure of f, a short computation yields

$$\sum_{\ell=0}^{k} f(\ell + m)\alpha_\ell = 0, \qquad m = 0, 1, \ldots. \tag{1}$$

The homogenous Hankel system (1) provides the coefficients α_ℓ of the Prony polynomial $\Lambda(z)$, and the unknown parameters T_j can now be extracted from the zeros of $\Lambda(z)$. Afterwards, the coefficients c_j are obtained by solving a linear system.

In recent years, the Prony method has been successfully applied to different inverse problems as e.g. for analysis of ultrasonic signals or for the approximation of Green functions in quantum chemistry or fluid dynamics, see e.g. [2, 3]. The renaissance of Prony's method originates from some modifications of the corresponding algorithm that considerably stabilize the original approach, [4, 7].

Searching the literature, one finds different further reconstruction methods that are closely related to Prony's method at second glance. In spectral analysis the

annihilating filter method is frequently applied. This idea has also been used already long ago for the construction of cyclic codes, [8]. For a given signal $S[n]$, the FIR filter $A[n]$ is called annihilating filter of $S[n]$, if

$$(A * S)(n) = \sum_{j\in\mathbb{Z}} A[j]\, S[n-j] = 0.$$

Using the z-transform $A(z) = \sum_{n=0}^{k} A[n]z^{-n}$ and comparing this equation to (1), we observe that $z^k A(z)$ undertakes the task of the Prony-polynomial.

In computer algebra, one is faced with the computation and processing of multivariate polynomials of high order. But if the polynomial f is k-sparse, i.e.,

$$f(x_1,\dots,x_n) = \sum_{j=1}^{k} c_j x_1^{d_{j_1}} x_2^{d_{j_2}} \cdots x_n^{d_{j_n}}$$

with $c_1,\dots,c_k \in \mathbb{C}$ and with k pairwise different vectors $(d_{j_1},\dots,d_{j_n}) \in \mathbb{N}^n$, the polynomial can be completely recovered using only $2k$ suitably chosen function values. Here again, the number of needed evaluations does not depend on the degree of the polynomial f but on the number k of active terms. The corresponding algorithm goes back to Ben-Or and Tiwari [1], and has recently been shown to be closely related to the Prony method. In [6], we considered the function reconstruction problem for sparse Legendre expansions of order N of the form

$$f(x) = \sum_{j=1}^{k} c_j P_{n_j}(x)$$

with $0 \le n_1 < n_2 \ldots < n_k = N$, where $k \ll N$, aiming at a generalization of Prony's method for this case. We succeeded to derive a reconstruction algorithm involving the function and derivative values $f^{(\ell)}(1)$, $\ell = 0,\dots,2k-1$. The reconstruction is based on special properties of Legendre polynomials and does not provide an idea for further generalization of the method to other orthogonal polynomial bases or to other function systems apart from exponentials and monomials.

Just recently, we developed a new perception of Prony's method based on eigenfunctions of linear operators, see [5]. This new insight gives us a tool for unification of all Prony-like methods on the one hand and for an essential generalization of the Prony approach on the other hand. This generalization will open a much broader field of applications of the method.

References

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

[2] G. Beylkin, L. Monzón, *Approximation by exponential functions revisited*, Appl. Comput. Harmon. Anal. **28** (2010), 131–149.

[3] F. Boßmann, G. Plonka, T. Peter, O. Nemitz and T. Schmitte, *Sparse deconvolution methods for ultrasonic NDT*, Journal of Nondestructive Evaluation (2012), to appear.

[4] Y. Hua and T.K. Sarkar, *Matrix pencil method for estimating parameters of exponentially damped/undamped sinusoids in noise*, IEEE Trans. Acoust. Speech Signal Process. **38**(5) (1990), 814–824.
[5] T. Peter and G. Plonka, *A generalized Prony method for reconstruction of sparse sums of eigenfunctions of linear operators*, University of Göttingen, preprint (2012).
[6] T. Peter, G. Plonka and D. Roşca, *Representation of sparse Legendre expansions*, J. Symbolic Comput., to appear (2012).
[7] D. Potts and M. Tasche, *Nonlinear approximation by sums of nonincreasing exponentials*, Appl. Anal. **90** (2011), 1631–1642.
[8] P. Stoica, R, Moses, *Introduction to Spectral Analysis*, Prentice Hall, Englewood Cliffs, NJ (2000).

Principal components of cumulants

Lek-Heng Lim
(joint work with Jason Morton)

Gian-Carlo Rota famously said that "Even today, the statistical theory of cumulants wears a halo of mystery that we still are a long way from dispelling. We do not hesitate to predict that cumulants will soon be inserted in the mainstream of mathematics." That was in 1986 and Rota's prediction did not materialize — cumulants are still as mysterious as they were a quarter century ago.

We would like to propose an explanation for this: Too much has been focussed on the combinatorics of cumulants and too little on its geometry. In this talk, we would like to discuss the geometry underlying cumulants and examine two unusual ways to analyze cumulants akin to principal components analysis: (i) decomposing a homogeneous form into a linear combination of powers of linear forms; (ii) decomposing a symmetric tensor into a multilinear combination of points on a Stiefel manifold. In the latter, one may identify 'principal cumulant components' via optimization over a Grassmannian.

Why might such principal components be useful? Multivariate Gaussian data are completely characterized by their mean and covariance but higher-order cumulants are unavoidable in non-Gaussian data. For univariate data, these are well-studied via skewness and kurtosis but for multivariate data, these cumulants are tensor-valued — higher-order analogs of the covariance matrix capturing higher-order dependence in the data. We argue that multivariate cumulants may be studied via these principal components, defined in a manner analogous to the usual principal components of a covariance matrix. It is best viewed as a subspace selection method that accounts for higher-order dependence the way PCA obtains varimax subspaces. A variant of stochastic gradient descent on the Grassmannian permits us to estimate principal components of cumulants of any order in excess of $10,000$ dimensions readily on a laptop computer.

A regularized spectral algorithm for Hidden Markov Models

Hà Quang Minh
(joint work with Marco Cristani, Alessandro Perina, Vittorio Murino)

Hidden Markov Models (HMM) are among the most important and widely used techniques in statistical learning, with numerous applications in various domains involving sequence modeling, including speech recognition, computer vision and pattern recognition, and bioinformatics. Traditionally, algorithms for learning HMMs have mainly employed Expectation Minimization (EM) [1]. While powerful and widely used, the main problem of EM methods is that they are prone to local minima. The quest for algorithms which are free of local minima and which are statistically consistent has been the focus of much research in the last decade. Two recent generalizations of HMMs, which are closely related, are Observable Operator Models (OMMs) [3] and Predictive State Representations (PSRs) [5]. Instead of the structure of unknown hidden states and emission probabilities, these models focus entirely on observation quantities and express sequence trajectories using linear operators, thus transforming probabilistic problems into linear algebraic ones.

Two recent algorithms implementing OMMs are [2, 4]. The main problem of these algorithms is that they are not very stable numerically. While they return exact results on exact observation statistics, on *empirical* observation statistics, which are what we have in practice, they often return probabilities which are negative or greater than one. This is due to their use of the Singular Value Decomposition and the pseudo-inverse operations.

Our contributions

To overcome numerical instability, we propose a regularized spectral algorithm. Specifically, let n be the number of possible symbols emitted by the HMM. Consider $P_1 \in \mathbb{R}^n$, $P_{2,1} \in \mathbb{R}^{n\times n}$, $P_{3,x,1} \in \mathbb{R}^{n\times n}$, which are defined by:

$$(P_1)_i = \mathbb{P}(x_1 = i), \tag{1}$$

$$(P_{2,1})_{ij} = \mathbb{P}(x_2 = i, x_1 = j), \tag{2}$$

$$(P_{3,x,1})_{ij} = \mathbb{P}(x_3 = i, x_2 = x, x_1 = j), \tag{3}$$

for $1 \leq x \leq n$. We approximate the probability

$$\mathbb{P}(X_1 = x_1, \ldots, X_t = x_t) \tag{4}$$

by a sequence of matrix multiplications

$$b_\infty^T B_{x_t} \ldots B_{x_1} b_1, \tag{5}$$

where

$$b_\infty = (U^T P_{2,1} P_{2,1}^T U + \gamma I)^{-1} (U^T P_{2,1} P_1), \tag{6}$$

$$B_x = (U^T P_{3,x,1} P_{2,1}^T U)(U^T P_{2,1} P_{2,1}^T U + \gamma I)^{-1}, \tag{7}$$

$$b_1 = U^T P_1, \tag{8}$$

for some regularization parameter $\gamma > 0$. Here U is a randomly chosen matrix of size $n \times k$, such that $rank(U^T P_{2,1} P_{2,1}^T U) = rank(P_{2,1}) = k$.

Compared to [2, 4], our algorithm

(1) is guaranteed to produce probability values that are always physically meaningful, that is between 0 and 1;
(2) on synthetic mathematical models, produces probability values that approximate very well true theoretical values;
(3) places no restriction on the number of symbols and the number of states. The theoretical justification for this case is significantly different from the case the number of states is smaller than or equal to the number of symbols.

Our algorithm has been tested on various real data sets in pattern recognition, showing significant improvements over classical HMMs, in both accuracy and speed.

References

[1] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, *A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains*, Annals of Mathematical Statistics **41**(1) (1970), 164–171.

[2] D. Hsu, S. Kakade, and T. Zhang, *A spectral algorithm for learning Hidden Markov models*, Proc. Conf. Learning Theory (COLT), (2009).

[3] H. Jaeger, *Observable operator models for discrete stochastic time series*, Neural Computation **12**(6) (2000), 1371–1398.

[4] S. M. Siddiqi, B. Boots, and G. Gordon, *Reduced-rank Hidden Markov models*, Proc. Intern. Conf. Artificial Intelligence and Statistics (AISTATS), (2010).

[5] S. Singh, M. James, and M. Rudary. *Predictive state representations: A new theory for modeling dynamical systems.* Proc. Conf. Uncertainty in Artificial Intelligence (UAI), (2004).

Can we estimate the density from an unweighted, random k-nearest neighbor graph?

Ulrike von Luxburg

Setting. Assume we draw a set of points $X_1, ..., X_n$ i.i.d. from some nice density p on $\mathbf{R}^d$. We build the k-nearest neighbor graph on this sample: its vertices correspond to the data points and X_i is connected to X_j by an undirected, unweighted edge whenever X_i is among the k nearest neighbors of X_j. Let A be the adjacency matrix of the graph.

The problem. Consider the following open problems:

- Can we estimate the density $p(X_i)$ at the data points (up to constant factors), just by looking at the adjacency matrix A?
- Can we approximately reconstruct the point locations $X_1, ..., X_n$ (up to translation, rotation, rescaling) when we just know the adjacency matrix A and n is large enough?

I believe that an answer to this problem is important for machine learning. Graph learning algorithms like spectral clustering or semi-supervised learning are often applied to unweighted kNN-graphs. If it turned out that the adjacency matrix of this graph is not informative enough about the underlying density, then we could not expect the machine learning algorithms to find the correct answers.

Techniques. The two problems stated above are closely related to each other in the sense that a solution to one of them implies a solution to the other one. In my talk I am going to discuss various, very diverse ideas and approaches to tackle the problem, but won't give final answers yet.

Just to make everybody curious, here are a couple of keywords that are relevant to the question:

- Mathematical geometry: sphere-preserving maps and Möbius maps
- Computational geometry: arrangements of hyperplanes
- Statistics: non-metric multidimensional scaling; ranking
- Analysis: completely monotonic functions

Answers? I don't have firm answers yet, just a couple of conjectures. Perhaps, the joint expertise of the participants would help to crack the nut!

Function approximation on data defined manifolds

Hrushikesh Mhaskar

We give a brief survey of our recent work on approximation of functions of unstructured, high dimensional data sets. One can assume that the data set is a sample from an unknown low dimensional manifold. While diffusion maps have been used to study the geometry of this manifold, we have developed a theory of wavelet-like representation of functions on the unknown manifold based on the eigenfunctions of the heat kernel. In turn, the heat kernel can be approximated using the data set, as is well knwon from the theory of Laplacian eigenmaps and diffusion maps. We describe also some applications of this theory to the analysis of some practical data sets.

Statistical inference in compound functional models

Alexandre B. Tsybakov
(joint work with Arnak S. Dalalyan, Yuri Ingster)

Assume that we observe a real-valued Gaussian process $\boldsymbol{Y} = \{Y(\phi)\colon \phi \in L^2([0,1]^d)\}$ such that

$$\mathbf{E}_f[Y(\phi)] = \int_{[0,1]^d} f(\boldsymbol{x})\,\phi(\boldsymbol{x})\,d\boldsymbol{x}, \quad \mathbf{Cov}_f(Y(\phi), Y(\phi')) = \varepsilon^2 \int_{[0,1]^d} \phi(\boldsymbol{x})\phi'(\boldsymbol{x})\,d\boldsymbol{x},$$

for all $\phi, \phi' \in L^2([0,1]^d)$, where $\mathbf{E}_f$ and $\mathbf{Cov}_f$ are the expectation and covariance signs and ε is some positive number. Our aim is to estimate the unknown function $f \in L^2([0,1]^d)$.

We denote by $L_0^2([0,1]^d)$ the subset of $L^2([0,1]^d)$ containing all the functions f such that $\int_{[0,1]^d} f(\boldsymbol{x})\,d\boldsymbol{x} = 0$. Let $\|\cdot\|_2$ denote the $L^2([0,1]^d)$-norm. For every $s \in \{1,\dots,d\}$ and $m \in \mathbb{N}$, we define $\mathcal{V}_s^d = \big\{V \subseteq \{1,\dots,d\} : |V| \le s\big\}$ and $\mathcal{B}_{s,m}^d = \big\{\boldsymbol{\eta} \in \{0,1\}^{\mathcal{V}_s^d} : |\boldsymbol{\eta}|_1 = m\big\}$ where $|V|$ is the cardinality of V and $|\boldsymbol{\eta}|_1$ is the number of ones in $\boldsymbol{\eta}$. For a vector $\boldsymbol{v}$, we denote by $\mathrm{supp}(\boldsymbol{v})$ its support, that is the set of indices of its non-zero components.

In order to circumvent the curse of dimensionality when estimating f, it is necessary to impose some a priori *structural* conditions on f. In the present work, we introduce a new type of structural assumption that has the following form.

Compound functional model. *There exists an integer $s \in \{1,\dots,d\}$, a binary sequence $\boldsymbol{\eta} \in \mathcal{B}_{s,m}^d$ and a set of functions $\{f_V \in L_0^2([0,1]^{|V|})\}_{V\in\mathcal{V}_s^d}$ such that*

$$f(\boldsymbol{x}) = \bar{f} + \sum_{V\in\mathcal{V}_s^d} f_V(\boldsymbol{x}_V)\eta_V = \bar{f} + \sum_{V\in\mathrm{supp}(\boldsymbol{\eta})} f_V(\boldsymbol{x}_V), \qquad \forall \boldsymbol{x} \in \mathbb{R}^d,$$

where $\bar{f} = \int_{[0,1]^d} f(\boldsymbol{x})\,d\boldsymbol{x}$.

In addition to this structural condition, we will also assume that the components f_V are smooth. Thus, given a collection $\boldsymbol{\Sigma} = \{\Sigma_V\}_{V\in\mathcal{V}_s^d}$ of subsets of $L_0^2([0,1]^s)$, we define the classes

$$\mathcal{F}_{s,m}(\boldsymbol{\Sigma}) = \bigcup\nolimits_{\boldsymbol{\eta}\in\tilde{\mathcal{B}}} \mathcal{F}_{\boldsymbol{\eta}}(\boldsymbol{\Sigma}),$$

where $\tilde{\mathcal{B}}$ is a given subset of $\mathcal{B}_{s,m}^d$ and

$$\mathcal{F}_{\boldsymbol{\eta}}(\boldsymbol{\Sigma}) = \Big\{f : \mathbb{R}^d \to \mathbb{R} : \exists\{f_V \in \Sigma_V\} \text{ such that } f = \bar{f} + \sum_V f_V\eta_V\Big\}.$$

The compound model is described by three main parameters. These are the dimension m that we call the *macroscopic* parameter, which characterizes the complexity of possible structure vectors $\boldsymbol{\eta}$, the dimension s that we call the *microscopic* parameter, which is responsible for the complexity of individual functions in the compound, and the complexity of functional class $\boldsymbol{\Sigma}$. The latter can be described by entropy numbers of $\boldsymbol{\Sigma}$ in convenient norms. We consider the case of Sobolev classes, $\Sigma_V = W_V(\beta, L)$ characterized by the smoothness $\beta > 0$ and the radius $L > 0$. We define the Sobolev class $W_V(\beta, L)$ by

$$W_V(\beta,L) = \Big\{g \in L_0^2([0,1]^d) :\ g = \sum_{\boldsymbol{j}\in\mathbb{Z}^d:\mathrm{supp}(\boldsymbol{j})\subseteq V} \theta_{\boldsymbol{j}}[g]\varphi_{\boldsymbol{j}} \text{ and } \sum_{\boldsymbol{j}\in\mathbb{Z}^d} |\boldsymbol{j}|_\infty^{2\beta}\theta_{\boldsymbol{j}}[g]^2 \le L\Big\}$$

where $\{\varphi_{\boldsymbol{j}}\}_{\boldsymbol{j}\in\mathbb{Z}^d}$ is a system of functions satisfying the appropriate conditions (for example, it can be the tensor-product trigonometric basis), $\theta_{\boldsymbol{j}}[f] = \langle f, \varphi_{\boldsymbol{j}}\rangle$ where $\langle\cdot,\cdot\rangle$ denotes the inner product in $L^2([0,1]^d)$, and $|\boldsymbol{j}|_\infty$ denotes the ℓ_∞ norm of $\boldsymbol{j} \in \mathbb{Z}^d$.

The integers m and s are "effective dimension" parameters. As soon as they grow, the structure becomes less pronounced and the compound model approaches the global nonparametric regression in dimension d, which is known to suffer from the curse of dimensionality already for moderate d. Therefore, an interesting case is the sparsity scenario where s and/or m are small.

We establish non-asymptotic upper and lower bounds on the minimax risk

$$\inf_{\widehat{f}} \sup_{f \in \mathcal{F}_{s,m}(\boldsymbol{W}(\beta,L))} \mathbf{E}_f[\|\widehat{f} - f\|_2^2],$$

where $\inf_{\widehat{f}}$ denotes the minimum over all estimators, and

$$\boldsymbol{W}(\beta, L) = \{W_V(\beta, L)\}_{V \in \mathcal{V}_s^d} .$$

We prove that, up to a multiplicative constant, the minimax risk behaves itself as

$$\max\left\{mL^{s/(2\beta+s)}\varepsilon^{4\beta/(2\beta+s)},\ ms\varepsilon^2 \log\left(\frac{d}{sm^{1/s}}\right)\right\} \tag{1}$$

(we assume here $d/(sm^{1/s}) > 1$, otherwise a constant factor should be inserted under the logarithm, see below). For the particular case $s = 1$, *i.e.*, for the additive regression model, [1] provides a lower bound on the minimax risk that matches (1) but an upper bound that departs from the lower one by a logarithmic factor. That paper assumes β to be known. We demonstrate that the rate (1) can be achieved for general s and in an adaptive way, that is without the knowledge of β, s, and m.

If $m = 1$, *i.e.*, $f(\boldsymbol{x}) = f_V(\boldsymbol{x}_V)$ for some unknown $V \subseteq \{1, \dots, d\}$ with $|V| \le s$, then the compound model reduces to a *single atom model*. For $s \ll d$, this can be viewed as a nonparametric generalization of the high-dimensional linear regression model under the sparsity scenario. The minimax rate of convergence (1) is then

$$\max\left\{L^{s/(2\beta+s)}\varepsilon^{4\beta/(2\beta+s)},\ s\varepsilon^2 \log\left(\frac{d}{s}\right)\right\}. \tag{2}$$

These rates account for two effects, namely, the accuracy of nonparametric estimation of f for fixed macroscopic structure parameter $\boldsymbol{\eta}$, cf. the first term $\sim \varepsilon^{4\beta/(2\beta+s)}$, and the complexity of the structure itself (irrespective to the nonparametric nature of the atoms $f_V(\boldsymbol{x}_V)$). In particular, the second term $\sim s\varepsilon^2 \log(d/s)$ in (2) coincides with the optimal rate of prediction in linear regression model under the standard sparsity assumption. It is important to note that the optimal rates depend only logarithmically on the ambient dimension d. Thus, even if d is large, the rate optimal estimators achieve nice performance under the sparsity scenario when s and m are small.

References

[1] G. Raskutti, M. Wainwright, B. Yu, *Minimax-optimal rates for sparse additive models over kernel classes via convex programming*, J. Mach. Learn. Res. **13** (2012), 389–427.

On approximations of the finite sample distribution of Support Vector Machines

Andreas Christmann
(joint work with Matías Salibían-Barrera, Stefan Van Aelst, Robert Hable)

The finite sample distribution of many nonparametric methods from statistical learning theory is unknown because the distribution P from which the data were generated is unknown and because there are often only asymptotical results on the behaviour of such methods known. The talk is focussed on the question how to draw statistical decisions (like confidence regions, prediction intervals, tolerance intervals or statistical tests) based on such nonparametric methods.

The first goal of this talk is to show that bootstrap approximations [8] of an estimator which is based on a continuous operator from the set of Borel probability distributions defined on a compact metric space into a complete separable metric space is qualitatively robust in the sense of robust statistics. As a special case it is shown that bootstrap approximations of (general) support vector machines (SVM) based on a Lipschitz continuous shifted loss function and a bounded kernel are qualitatively robust, both for the risk functional and for the SVM operator itself, if some relatively mild conditions are satisfied. Details of our results are given in [2] and can be interpreted as generalizations of theorems derived by [6].

The second goal of this talk is to show that bootstrap approximations of qualitatively robust support vector machines converge in outer probability, if some weak assumptions are satisfied. This result is unpublished.

(General) support vector machines based on some shifted loss function $L^\star$ and some reproducing kernel Hilbert space (RKHS) H – which is often specified via the corresponding kernel k – are defined by

$$f_{L^\star,\mathrm{P},\lambda} := \arg\inf_{f\in H} \mathbb{E}_\mathrm{P} L^\star(X,Y,f(X)) + \lambda\|f\|_H^2, \tag{1}$$

where P denotes the unknown distribution and $L : \mathcal{X}\times\mathcal{Y}\times\mathbb{R}\to[0,\infty)$ is convex with respect to its third argument. The shifted version of a loss function is given by $L^\star(x,y,t) := L(x,y,t) - L(x,y,0)$ for all $(x,y)\in\mathcal{X}\times\mathcal{Y}$. Given a data set $D = ((x_1,y_1),\ldots,(x_n,y_n)) \in (\mathcal{X}\times\mathcal{Y})^n$, the empirical SVM is of course defined by $f_{L^\star,\mathrm{D},\lambda}$, where D denotes the empirical distribution defined by D.

This class of statistical methods based on minimizing a regularized risk with the special regularizing term $\lambda\|f\|_H^2$ over an RKHS plays an important role in statistical machine learning. The original SVM approach by [1] was derived from the *generalized portrait algorithm* invented earlier by [15]. Considering regularized empirical (least squares) risks over reproducing kernel Hilbert spaces is a relatively old idea, see, e.g., [10] and [16] and the references therein.

Obviously, the definition of (general) SVMs given in (1) covers many specific loss functions as special cases, e.g., the hinge loss function for binary classification, the ϵ-insensitive loss function for regression, the check loss function (which is also called pinball loss function) for quantile regression, and the logistic loss functions

for classification and for regression. Note that these five loss functions are even Lipschitz continuous, and many authors have shown that the combination of a Lipschitz continuous loss function and a bounded and continuous kernel (e.g., a Gaussian RBF kernel or a Wendland kernel, see [17]) yields statistically robust (general) SVMs for classification and for regression purposes, see e.g. [9], [3], [11], and [4]. In general, this is not true for (general) SVMs based on a non-Lipschitz continuous loss function, if the output space $\mathcal{Y}$ is unbounded.

References

[1] B. E. Boser, I. Guyon, and V. Vapnik, *A training algorithm for optimal margin classifiers*, in: Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory, pp. 144–152, ACM, Madison, WI. (1992).

[2] A. Christmann, M. Salibían-Barrera, and S. Van Aelst, *On the stability of bootstrap estimators*, Preprint, `http://arxiv.org/abs/1111.1876`.

[3] A. Christmann and I. Steinwart, *Consistency and robustness of kernel based regression*, Bernoulli **13** (2007), 799–819.

[4] A. Christmann, A. Van Messem, and I. Steinwart, *On consistency and robustness properties of support vector machines for heavy-tailed distributions*, Statistics and Its Interface **2** (2009), 311–327.

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

[6] A. Cuevas and R. Romo, *On robustness properties of bootstrap approximations*, J. Statist. Plann. Inference **2** (1993), 181–191.

[7] L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*, Springer, New York (1996).

[8] B. Efron, *Bootstrap methods: another look at the jackknife*, Annals of Statistics **7** (1979), 1–26.

[9] S. Mukherjee, P. Niyogi, T. Poggio, and R. Rifkin, *Learning theory: stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization*, Adv. Comput. Math. **25** (2006), 161–193.

[10] T. Poggio and F. Girosi, *A theory of networks for approximation and learning*, Proc. IEEE **78** (1990), 1481–1497.

[11] I. Steinwart and A. Christmann, *Support Vector Machines*, Springer, New York (2008).

[12] J. A. K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, and J. Vandewalle, *Least Squares Support Vector Machines*, World Scientific, Singapore (2002).

[13] V. N. Vapnik, *The Nature of Statistical Learning Theory*, Springer, New York (1995).

[14] V. N. Vapnik, *Statistical Learning Theory*, John Wiley & Sons, New York (1998).

[15] V. N. Vapnik and A. Lerner, *Pattern recognition using generalized portrait method*, Autom. Remote Control **24** (1963), 774–780.

[16] G. Wahba, *Spline Models for Observational Data*, Series in Applied Mathematics, vol. 59, SIAM, Philadelphia (1990).

[17] H. Wendland, *Scattered Data Approximation*, Cambridge University Press, Cambridge (2005).

Nonuniform support recovery from noisy random measurements by orthogonal matching pursuit

Song Li

This talk considers nonuniform support recovery via Orthogonal Matching Pursuit (OMP) from noisy random measurements. Given m admissible random measurements (of which subgaussian measurements is a special case) of a fixed s-sparse signal x in $\mathbb{R}^n$ corrupted with additive noise, we show that under a condition on the minimum magnitude of the nonzero components of x, OMP can recover the support of x exactly after s iterations with overwhelming probability provided that $m = \mathcal{O}(s \log n)$. This extends the results of J. A. Tropp and A. C. Gillert to the case with noise. It is a real improvement over previous results in the noisy case, which are based on mutual incoherence property or restricted isometry property analysis and which require $\mathcal{O}(s^2 \log n)$ random measurements.

In addition, this talk also considers sparse recovery from noisy random frequency measurements via OMP. Similar results can be obtained for the partial random Fourier matrix via OMP provided that $m = \mathcal{O}(s(s + \log(ns)))$. Thus, for some special cases, this answers the open question raised by J .A. Tropp and H. Rauhut.

Learning in variable RKHSs with application to the blood glucose reading

Sergei V. Pereverzyev
(joint work with V. Naumova, S. Sivananthan)

In this talk we consider the problem of a reconstruction of a real-valued function $f : X \to \mathbb{R}, X \subset \mathbb{R}^d$, from a given data set

$$\mathbf{z} = \{(x_i, y_i)\}_{i=1}^n \subset X \times \mathbb{R},$$

where it is assumed that $y_i = f(x_i) + \xi_i$, and $\xi_i = \{\xi_i\}_{i=1}^n$ is a noise vector. The reconstruction problem can be considered in two aspects: (i) interpolation – to evaluate the value of a function $f(x)$ for $x \in \overline{co\{x_i\}}$, (ii) extrapolation – to predict the value of $f(x)$ for $x \notin \overline{co\{x_i\}}$.

In both aspects the reconstruction problem is ill-posed and one of the classical ways to solve it is the use of a Tikhonov-type method, which in the present context consists in constructing a regularized solution $f_\lambda(x)$ as a minimizer of the functional

$$T_\lambda(f; \mathcal{H}, \mathbf{z}) = \frac{1}{|\mathbf{z}|} \sum_{i=1}^{|\mathbf{z}|} (y_i - f(x_i))^2 + \lambda ||f||_{\mathcal{H}}^2, \tag{1}$$

where $|\mathbf{z}|$ is the cardinality of the set $\mathbf{z}$, i.e., $|\mathbf{z}| = n$, and λ is a regularization parameter.

The Tikhonov method (1) raises two main issues that should be clarified before use of this scheme. One of them is how to choose a regularization parameter λ. This problem has been extensively discussed [1, 2, 3, 5, 11]. Another one, which

is even more important, is how to choose the space $\mathcal{H}$, whose norm is used for a penalization. Despite its significance, the second issue is much less studied. Note there are still no general principles to advise a choice and only in a few papers [2, 8, 9, 12] some methods for finding an appropriate space for the regularization of ill-posed problems have been proposed. Keeping in mind that a Sobolev space $\mathcal{H} = W_2^r$ traditionally used in (1) is a particular example of a Reproducing Kernel Hilbert Space (RKHS), the above mentioned issue is about the choice of a kernel K for an RKHS $\mathcal{H} = \mathcal{H}_K$.

In this talk we describe a novel approach [10], so-called kernel adaptive regularized (KAR) algorithm, where the choice of the kernel and regularization parameter is governed by several conditions, which allow to achieve a good performance within the regularization procedure. In short, the proposed approach is based on splitting of a given data set $\mathbf{z}$ and is oriented towards extrapolation.

To be more specific, the kernel K is chosen from the set of admissible kernels $\mathcal{K}$ as the minimizer of the following functional

(2)
$$Q_\mu(K, \lambda(K), \mathbf{z}) = \mu T_{\lambda(K)}(f_{\lambda(K)}(\cdot; K, \mathbf{z}_T), \mathcal{H}_K, \mathbf{z}_T) + (1-\mu)P(f_{\lambda(K)}(\cdot; K, \mathbf{z}_T), K, \mathbf{z}_P),$$

where

$$\mathbf{z}_T \cup \mathbf{z}_P = \mathbf{z}, \ \overline{co\{x_i : (x_i, y_i) \in \mathbf{z}_T\}} \cap \{x_i : (x_i, y_i) \in \mathbf{z}_P\} = \emptyset,$$

$\mu \in [0, 1]$ is the parameter that can be seen as a performance regulator on the sets $\mathbf{z}_P$ and $\mathbf{z}_T$, and $f_\lambda(\cdot; K, \mathbf{z}_T)$ is the minimizer of the Tikhonov-type functional $T_\lambda(f; \mathcal{H}_K, \mathbf{z}_T)$.

At the same time, the functional P is used to measure the performance of the regularization estimator $f_\lambda(x; K, \mathbf{z}_T)$, constructed with the use of the data set $\mathbf{z}_T$, on the rest of a given data $\mathbf{z}_P$.

For a rather general form of the set $\mathcal{K}$ and parameter choice rule $\lambda = \lambda(K)$ we justify the existence of the kernel $K^0 \in \mathcal{K}$ that minimizes the functional (2).

The last part of the talk is concerned with the practical application of the proposed approach. Namely, we consider how the approach based on the minimization of the functional (2) can be adapted to a problem of diabetes therapy management, to be more specific, reading blood glucose (BG) concentration of diabetic patients from electrical signals in the interstitial fluid (ISF), measured by commercially available devices, Continuous Glucose Monitoring (CGM) systems. We illustrate the results of the numerical experiments with real clinical data and show advantages of this new approach, by comparing the performance of the constructed blood glucose estimators with the performance of the commercially available CGM-systems.

Finally, we discuss the versatility and effectiveness of the proposed approach for other applications from diabetes therapy management.

References

[1] F. Cucker, S. Smale, *On the mathematical foundations of learning*, Bull. Amer.Math. Soc. (N.S.) **39** (2009), 1–49.

[2] E. De Vito, S. Pereverzyev, L. Rosasco, *Adaptive kernel methods using the balancing principle*, Found. Comput. Math. **10** (2010), 455-479.
[3] H. Engl, M. Hanke, A. Neubauer, *Regularization of Inverse Problems*, volume 375 of Mathematics and Its Applications, Kluwer Academic Publishers, Dordrecht, Boston, London, 1996.
[4] A. Facchinetti, G. Sparacino, C. Cobelli, *Reconstruction of glucose in plasma from interstitial fluid continuous glucose monitoring data: role of sensor calibration*, J. Diabetes Sci. Technol. **1** (2007), 617–623.
[5] S. Kindermann, A. Neubauer, *On the convergence of the quasi-optimality criterion for (iterated) Tikhonov regularization*, Inverse Problems and Imaging **2** (2008), 291–299.
[6] D. W. Klonoff, *Continuous glucose monitoring: roadmap for 21-st diabetes therapy*, Diabetes Care **28** (2005), 1231–1239.
[7] B. Kovatchev, W. Clarke, *Peculiarities of the continuous glucose monitoring data stream and their impact on developing closed-loop control technology*, J. Diabetes Sci Technol. **2** (2008), 158–163.
[8] G. R. G. Lanckriet, N. Christianini, L.E. Ghaoui, P. Bartlett, M.I. Jordan, *Learning the kernel matrix with semidefinite programming*, J. Mach. Learn. Res. **5** (2004), 27–72.
[9] C. A. Micchelli, M. Pontil, *Learning the kernel function via regularization*, J. Mach. Learn. Res. **6** (2005), 1099-1125.
[10] V. Naumova, S. V. Pereverzyev, S. Sampath, *Extrapolation in variable RKHSs with application to the blood glucose reading*, Inverse Problems **27** (2011), 075010, 13 pp.
[11] A. N. Tikhonov, V. B. Glasko, *Use of the regularization methods in non-linear problems*, USSR Comput. Math. Phys. **5** (1965), 93–107.
[12] G. Wahba, *Spline Models for Observational Data*, volume 59 of Series in Applied Mathematics, CBMS-NSF Regional Conf., SIAM (1990).

Eigenvalues, entropy numbers, and Mercer representations for RKHSs

Ingo Steinwart

Reproducing kernel Hilbert spaces (RKHSs) play an important role in many modern machine learning methods including both supervised methods such as support vector machines (SVMs) and related regularized kernel methods, and unsupervised methods such as kernel PCA and some manifold techniques. In the analysis of these learning algorithms one frequently needs to quantify the approximation properties of these spaces, e.g. in terms of eigenvalues, entropy numbers, or interpolation spaces.

For example, one of the currently sharpest techniques to bound the estimation error of SVMs and related methods uses a localized ansatz together with Talagrand's inequality, symmetrization, peeling, and Dudley's entropy integral. Since the latter can be equivalently expressed in terms of entropy numbers, it turns out that expected entropy numbers of the form

$$\mathbb{E}_{D\sim\nu^n} e_i\big(\mathrm{id} : H \to L_2(D)\big)$$

need to be considered in order to bound the estimation error. Here ν is the marginal distribution of the data generating distribution P, e_i denotes the ith (dyadic) entropy number, H is the considered RKHS, and $L_2(D)$ is the Lebesgue space with respect to the empirical measure defined by the data $D = (x_1, \ldots, x_n)$.

For general Banach spaces, bounding such expected entropy numbers *directly* is known to be extremely difficult, and hence one usually resorts to bounding

$$\mathbb{E}_{D\sim\nu^n} e_i\big(\mathrm{id} : H \to \ell_\infty(X)\big)\,,$$

instead. The first part of the talk, which is based on [2], shows that for RKHSs this crude approach can often be improved. To be more precise, assume that we have a constant c such that

$$e_i\big(I_k : H \to L_2(\nu)\big) \le c\, i^{-\frac{1}{2p}}\,, \qquad i \ge 1\,, \tag{1}$$

where $I_k : H \to L_2(\nu)$ denotes the "inclusion" that maps an $f \in H$ to its equivalence class $[f]_\sim$ in $L_2(\nu)$. Then there exists another constant K_p only depending on p such that

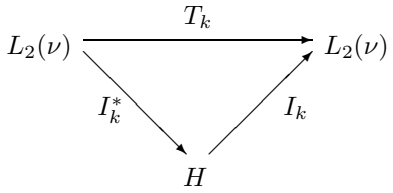
$$\mathbb{E}_{D\sim\nu^n} e_i\big(\mathrm{id} : H \to L_2(D)\big) \le K_p\, c\, i^{-\frac{1}{2p}}\,, \qquad i \ge 1,\, n \ge 1\,.$$

Moreover, (1) is shown to be equivalent to

$$\lambda_i(T_k) \le C\, i^{-\frac{1}{p}}\,, \qquad i \ge 1\,, \tag{2}$$

where C is some constant independent of i and $T_k : L_2(\nu) \to L_2(\nu)$ is the integral operator associated to the kernel k of the RKHS H. As a consequence of these results one can easily describe the "capacity" of the RKHS H in terms of the eigenvalues of T_k for most regularized kernel methods.

Apart from various tools of functional analysis, the proof relies on the decomposition



where the adjoint I_k^* of I_k is "integral" operator

$$\begin{aligned} S_k : L_2(\nu) &\to H \\ f &\mapsto \int_X k(x,\,\cdot\,) f(x)\, d\nu(x) \end{aligned}$$

This decomposition is also used in the second part of the talk, in which the classical Mercer series representation for continuous kernels on compact domains is extended to almost arbitrary kernels. To be more precise, Mercer's classical theorem yields the representation

$$k(x,x') = \sum_{i\in I} \lambda_i e_i(x) e_i(x')\,, \tag{3}$$

where λ_i is the ith eigenvalue of the integral operator T_k and (e_i) is a corresponding ONS of eigenfunctions. Here the convergence of the series above is both absolute and pointwise. It is well-known that with the help of (3) one can describe the

RKHS H of k, and in turn this description makes it possible to describe e.g. the approximation properties of H in terms of interpolation spaces.

Now, we have shown in [3] that, for separable RKHSs H, there exists ν-zero set $N \subset X$ such that

$$k(x, x') = \sum_i \lambda_i e_i(x) e_i(x') \,, \qquad x, x' \in X \setminus N,$$

where λ_i and e_i are as above. A first consequence of this result is that for separable RKHSs one can use a Mercer representation for ν^n-almost all Gram matrices $(k(x_i, x_j))_{i,j=1}^n$, which in one or the other form is the basis of most kernel-based learning algorithms. A second consequence is that the images of the fractional powers $T_k^{\beta/2}$ of T_k can be exactly described by the interpolation spaces

$$\big[L_2(\nu), [H]_\sim\big]_{\beta,2}$$

of the real method, where $[H]_\sim$ denotes the image of I_k in $L_2(\nu)$. The latter result extends and clarifies a similar description of $T_k^{\beta/2}$ in [1], which is important to describe the approximation properties of H.

References

[1] S. Smale and D.-X. Zhou, *Estimating the approximation error in learning theory*, Anal. Appl. **1** (2003), 17–41.

[2] I. Steinwart, *Oracle inequalities for SVMs that are based on random entropy numbers*, J. Complexity **25** (2009), 437–454.

[3] I. Steinwart and C. Scovel, *Mercer's theorem on general domains: on the interaction between measures, kernels, and RKHSs*, Constr. Approx. **35** (2012), 363–417.

Methods of kernel construction

Robert Schaback

Kernels provide an important link between Approximation and Learning Theory. For a survey on kernel *applications*, see [8]. This talk focuses on certain methods for explicit constructions of new positive definite kernels, in particular kernels that are linked to generalized Sobolev spaces.

1. Introduction

Since the talk has no time to consider kernels based on expansions and special kernels connected to Partial Differential Equations (see the survey [2] for some cases), it focuses on *radial* kernels on $\mathbb{R}^d$ (a.k.a. *radial basis functions*), i.e.

$$K \,:\, \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R},\ K(x,y) = \phi(r),\ r := \|x-y\|_2 \text{ for all } x, y \in \mathbb{R}^d.$$

If kernels are positive semidefinite, they are in one–to–one correspondence to "native" Hilbert spaces in which they are reproducing. In contrast to Machine Learning, where kernels arise from feature maps and the corresponding Hilbert spaces stay in the background, Numerical Analysis and Approximation Theory start with

Hilbert spaces and look for their reproducing kernels. In all cases, the construction of useful kernels is of extreme significance.

The most important Hilbert spaces for Numerical Analysis are Sobolev spaces $W_2^m(\mathbb{R}^d)$. Spaces on bounded Lipschitz domains can be handled by restriction (see 10.7 of [11]). The global versions can be defined via the weight function $(1+r^2)^m$ in the Fourier domain, and thus their reproducing kernels are the inverse d–variate Fourier transforms of $(1+r^2)^{-m}$, i.e. the Whittle–Matérn kernels

$$M_\nu(r) := r^\nu K_\nu(r),\ r \geq 0,\ \nu = m - d/2 \tag{1}$$

where the K_ν are the Bessel functions of second kind. The index ν can be real, and thus one can handle cases of fractional order and dimension.

We shall focus here on Hilbert spaces which are norm–equivalent to Sobolev spaces and construct their reproducing kernels.

2. Fractional derivatives of radial kernels

But beforehand we describe how to take fractional derivatives of radial kernels. This is particularly useful for programming purposes [6]. Using tools from [9] one can verify

Observation 1. *All standard classes of radial kernels are closed under fractional derivatives, tf their elements ϕ are written in f–*form $f(s) := \phi(\sqrt{2s})$.

The f–form is well–known from the correspondence between positive definite and completely monotone functions. The fractional derivatives considered here are real powers of the differential operator $D\ :\ f \mapsto -f'$, and the classes are of the form $\{f_\alpha := D^\alpha f_0\}_{\alpha \in A \subset \mathbb{R}}$ for some fixed function f_0. This makes ideal trial spaces for solving fractional differential equations. Furthermore, the identity

$$F_{d'}^{-1} \circ F_d = D^{(d'-d)/2}$$

holds if F_d is the radial Fourier transform in $\mathbb{R}^d$ (see [9], written without knowing Matheron's work [4]). Thus closedness under forward–backward Fourier transformation in different fractional dimensions is equivalent to closedness under fractional derivatives.

Simple examples are

- *Gaussians*: $\phi(r) = \exp(-r^2/2),\ f(s) = \exp(-s),\ D^\alpha f = f$,
- *Whittle–Matérn Kernels* (1): $f_\nu(s) = (\sqrt{2s})^\nu K_\nu(\sqrt{2s})$ with $D^\alpha f_\nu = f_{\nu-\alpha}$,
- *Wendland kernels* $\phi_{d,k}(r)$ with $f_{d,k}(s) = \phi_{d,k}(\sqrt{2s})$ and $D^\alpha f_{d,k} = f_{d+2\alpha,k-\alpha}$

for certain ranges of α, but, until recently, the Wendland functions $\phi_{d,k}(r)$ were only defined for integer d and k. We turn to this now.

3. Generalized Wendland functions

To generalize the compactly supported Wendland [10] functions $\phi_{d,k}$ for integer $d \geq 1$ and integer $k \geq 0$ which are positive definite in $\mathbb{R}^d$ and polynomials of degree $\lfloor d/2 \rfloor + 3k + 1$ and reproducing in Hilbert spaces norm–equivalent to

$W_2^{\lfloor d/2\rfloor+k+1/2}(\mathbb{R}^d)$, one can use the formula $\phi_{d,k}=\psi_{\lfloor d/2\rfloor+k+1,k}$ and the representation

$$\psi_{\mu,\alpha}(x)=\int_x^1 r(1-r)^\mu \frac{(r^2-x^2)_+^{\alpha-1}}{\Gamma(k)2^{\alpha-1}}dr \text{ for all } \mu\geq 0,\alpha>0,\ 0\leq x\leq 1 \tag{2}$$

which yields polynomials for integer μ and α. In [7], the cases generating the "missing" integer order Sobolev spaces were added, i.e. the functions $\phi_{d,k}$ with half–integer k. The first interesting case is $d=2,\ k=1/2$ with

$$\phi_{2,1/2}(x) \quad = \quad \frac{\sqrt{2}}{3\sqrt{\pi}}\left(3x^2\log\left(\frac{x}{1+\sqrt{1-x^2}}\right)+(2x^2+1)\sqrt{1-x^2}\right)$$

It is a reproducing kernel in a norm–equivalent space to $W_2^2(\mathbb{R}^2)$.

Theorem 1. [7]
All Wendland functions with k being a half–integer are linear combinations of even polynomials with factors $\log\left(\frac{x}{1+\sqrt{1-x^2}}\right)$ *and* $\sqrt{1-x^2}$.

The paper [7] conjectures that hypergeometric functions might provide the full solution for arbitrary $\mu\geq 0$ and $\alpha\geq 0$.

This problem was recently solved by Simon Hubbert [3] in terms of Associate Legendre functions

$$\begin{array}{rl} & P_\alpha^{-(\alpha+\mu)/2}(z) \\ = & \frac{1}{\Gamma(1+(\alpha+\mu)/2)}\left(\frac{1+z}{1-z}\right)^{-(\alpha+\mu)/4} {}_2F_1\left(-\alpha,\alpha+1;1+(\alpha+\mu)/2;(1-z)/2\right) \end{array}$$

as

$$\psi_{\mu,\alpha}(r)=\Gamma(\mu+1)(1-r^2)^{(\mu+\alpha)/2}r^\alpha P_\alpha^{-(\alpha+\mu)}\left(\frac{1}{r}\right),\ \mu>-1,\ \alpha>0.$$

By direct inspection of (2) one gets

Theorem 2. *In f–form* $f_{\mu,\alpha}(s)=\psi_{\mu,\alpha}(\sqrt{2s})$*, this class is closed under fractional derivatives:*

$$D^\beta f_{\mu,\alpha}=f_{\mu,\alpha-\beta}$$

as far as the application of the operators is well–defined.

4. Generalized Whittle–Matérn kernels

The Fourier weight of the classical Sobolev spaces can be slightly generalized to $(\kappa^2+\|\omega\|_2^2)^m$ to get the scaled version

$$\phi_\kappa(r)=\frac{2^{1-m}}{(m-1)!}\left(\frac{r}{\kappa}\right)^{m-d/2}=\frac{2^{1-m}}{(m-1)!}\kappa^{d-2m}M_{m-d/2}(\kappa r). \tag{3}$$

of (1) as a reproducing kernel of a space norm–equivalent to $W_2^m(\mathbb{R}^d)$. The corresponding Fourier transform is $(\kappa^2+\|\omega\|_2^2)^{-m}$, but a considerably more difficult

problem is the Fourier inversion of the radial function

$$\prod_{j=1}^{m}(\kappa_j^2 + r^2) \tag{4}$$

for different κ_j instead. This would yield a norm–equivalent Hilbert space to $W_2^m(\mathbb{R}^d)$ again, if all κ_j are positive. Using the divided difference $[\ldots]_z$ with respect to a variable z the result is

Theorem 3. [1] *If all κ_j are nonzero, the d–variate radial Fourier transform of (4) is*

$$\phi(r) = 2^{-m+1}(-1)^{m-1}[\kappa_1^2/2, \ldots, \kappa_m^2/2]_z \left(\frac{r}{\sqrt{2z}}\right)^{1-d/2} K_{1-d/2}(r\sqrt{2z}),$$

and it equals (3) for a variable scale $\kappa(r)$ between $\kappa_1, \ldots, \kappa_m$.

The proof uses the relation

$$(-1)^{m-1}\prod_{j=1}^{m}(s+t_j)^{-1} = [t_1, \ldots, t_m]_z(s+z)^{-1}$$

and relies heavily on the second section on fractional derivatives otherwise.

In case that $k > d/2$ of the κ_j vanish, the resulting kernel can be shown [1] to contain polyharmonic splines

$$\begin{cases} r^{2k-d} & d \text{ odd}, \\ r^{2k-d}\log r & d \text{ even}. \end{cases}$$

This forces to consider conditionally positive definite kernels and generalized Fourier transforms. The associated native Hilbert spaces should be some crossover between Sobolev and Beppo–Levi spaces, but are not yet investigated.

5. Open problem

The talk closes with my favourite problem in kernel construction:

> Find an *explicit* formula for a radial, positive definite, compactly supported and infinitely differentiable kernel.

Such kernels must exist by simple convolution arguments, and a good example would be the refinable up–function [5] if it were explicitly known. A multivariate analogon could result as the inverse Fourier transform of an infinite product of squares of Bessel functions. Any progress would be much appreciated.

References

[1] M. Bozzini, M. Rossini, and R. Schaback. Generalized Whittle–Matérn and polyharmonic kernels. to appear, 2012.

[2] St. De Marchi and R. Schaback. Nonstandard kernels and their applications. *Dolomites Research Notes on Approximations*, 2:16–43, 2009.

[3] S. Hubbert. Closed form representations for a class of compactly supported radial basis functions. *Advances in Computational Mathematics*, 36:115–136, 2012.

[4] G. Matheron. *Les variables régionaliseés et leur estimation.* Masson, Paris, 1965.

[5] V.A. Rvachev. Compactly supported solutions of functional-differential equations and their applications. *Russian Mathematical Survey*, 45:87–120, 1990.
[6] R. Schaback. Matlab programming for kernel-based methods. Technical report, available via http://num.math.uni-goettingen.de/schaback/research/papers/MPfKBM.pdf.
[7] R. Schaback. The missing Wendland functions. *Advances in Computational Mathematics*, 43:76–81, 2010.
[8] R. Schaback and H. Wendland. Kernel techniques: from machine learning to meshless methods. *Acta Numerica*, 15:543–639, 2006.
[9] R. Schaback and Z. Wu. Operators on radial basis functions. *J. Comp. Appl. Math.*, 73:257–270, 1996.
[10] H. Wendland. Piecewise polynomial, positive definite and compactly supported radial functions of minimal degree. *Advances in Computational Mathematics*, 4:389–396, 1995.
[11] H. Wendland. *Scattered Data Approximation*. Cambridge University Press, 2005.

Kernel based methods in Learning and Approximation

Wolfgang zu Castell
(joint work with Georg Berschneider)

Efficient learning builds upon the proper choice of an a priori model on the space of structures to be learned. This so-called inductive bias can be suitable modelled using kernels and their associated reproducing kernel spaces. The resulting mathematical model can then be framed within the context of regularization networks. Within this framework, the goal is to determine a hypothesis from a given hypothesis space H minimizing the emprical risk

$$R_e[f] = \frac{1}{N}\sum_{j=1}^{N} c(x_j, y_j, f(x_j)) \; + \; \lambda p(|f|_H),$$

where $\{(x_j, y_j) \in X \times Y : 1 \leq j \leq N\}$ is some given data, $c : X \times X \times Y \to \mathbb{R}_+$ a loss function and $p : \mathbb{R} \to \mathbb{R}$ a monotonically increasing penalty function.

If $K : X \times X \to L(W)$ is a positive definite, operator-valued kernel (cf. [2, 3]), where $L(W)$ denotes the space of bounded, linear operators on a Hilbert space W, and H is chosen to be the associated reproducing kernel Hilbert space of W-valued functions on X, then the so-called representer theorem guarantees that there is a solution of the optimization problem which can be represented in the from

$$f(x) \; = \; \sum_{j=1}^{N} K(x, x_j) w_j.$$

This is a more or less straight-forward consequence of the reproducing property and Pythagoras' Theorem.

If K is conditionally positive definite with respect to a given finite-dimensional space U of W-valued functions on X, there is an associated reproducing kernel Pontryagin space $\Pi = H \oplus U$ with maximal negative subspace U (cf. [1, 3]).

According to the decomposition of the Pontryagin space, every function $f \in \Pi$ can be written as $f = f_1 + f_2$, where $f_1 \in H$ and $f_2 \in U$. For the interesting case of the regularization problem the penalty is applied to the norm of the component f_1

(e.g., the smoothing spline problem). For this situation, there is again a representer theorem, i.e., there exists a solution of the form

$$f(x) \;=\; \sum_{j=1}^{N} K(x, x_j) w_j \;+\; u(x),$$

where u is an appropriate function in U, and

$$\sum_{j=1}^{N} \langle g(x_j), w_j \rangle_W \;=\; 0$$

for all $g \in U$.

Using the geometric structure of the Pontryagin space, the proof can be referred back to the Hilbert space version of the theorem.

The method of proof can further be extended to the infinite-dimensional case, i.e., where Π is a Kreĭn space.

An extended version of this contribution is given in [4].

References

[1] G. Berschneider, W. zu Castell, S.J. Schrödl, *Conditionally positive definite kernels and Pontryagin spaces*, in: M. Neamtu, L.L. Schumaker (ads.), Approximation Theory XII: San Antonio 2007, , pp. 27–37, Nashboro Press, Brentwood (2008).

[2] R.K. Beatson, W. zu Castell, S.J. Schrödl, *Kernel based methods for vector-valued data with correlated components*, SIAM J. Sci. Comp. **33**(4) (2011), 1975–1995.

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

[4] G. Berschneider, W. zu Castell, *The representer theorem for reproducing kernel spaces*, preprint.

Classification algorithms using adaptive partitioning

Wolfgang Dahmen

(joint work with Peter Binev, Albert Cohen, Ronald DeVore)

Algorithms for binary classification based on adaptive partitioning are formulated and analyzed for both their risk performance and their friendliness to numerical implementation, see [3]. The algorithms can be viewed as generating a set approximation to the Bayes set and thus fall into the general category of *set estimators*. A general theory is developed to analyze the risk performance of set estimators with the goal of guaranteeing performance with *high probability* rather than in *expectation*. Convergence rates in expectation can easily be derived from the given estimates in probability. The analysis decouples the approximation and estimation effects on the risk. The estimation errors are dealt with by introducing a new *modulus*. Its relevance and usefulness hinges among other things on certain functions of measurable sets bounding the deviation of the estimation error from its empirical counterpart with high probability.

A crucial tool for controlling the involved uniform deviations is Talagrand's concentration inequality, see e.g. [1]. Studying the relation of the modulus to margin conditions (see e.g. [4]) leads to concrete bounds for the estimation error. Furthermore, bounds are given for the approximation error (bias) based on the smoothness of the regression function and margin conditions. When these approximation results are combined with the estimation error bounds, an estimate of risk performance is obtained. A simple model selection is used to optimally balance the approximation and estimation error bounds. This general theory is then applied to algorithms based on adaptive partitioning. Results are formulated for the risk performance of these algorithms in terms of Besov smoothness of the regression function and margin conditions. Adaptivity allows one to relax classical Hölder smoothness (smoothness in L_∞) to weaker Besov smoothness (smoothness in L_p) [5]. In particular, this increases the compatibility range for margin conditions and the order of smoothness of the regression function.

The results of this paper are related to the work of Scott and Nowak [6] on tree based adaptive methods for classification, however, with several important distinctions. In particular, our model selection utilizes a validation sample to avoid identifying suitable penalty terms. This allows us to employ *wedge decorated* trees that yield higher order performance. Finally, it is briefly indicated how to accommodate plug-in estimators in this framework. The desired bounds with high probability would then follow from corresponding bounds in probability (rather than in expectation) for the regression function. However, such bounds for general measures do not hold for higher order piecewise polynomial estimators (see [2]) while the above approach does allow us to obtain higher order performance with high probability.

References

[1] P. Bartlett and S. Mendelson, *Empirical minimization*, Prob. Theory and Related Fields **135** (2003), 311–334.

[2] P. Binev, A, Cohen, W. Dahmen, and R. DeVore, *Universal algorithms for learning theory, part II: piecewise polynomials*, Constructive Approximation **26**(2007), 127–152.

[3] P. Binev, A, Cohen, W. Dahmen, and R. DeVore, *Classification algorithms using adaptive partitioning*, IGPM Report # 343, July 2012, RWTH Aachen.

[4] Olivier Bousquet, Stéphane Boucheron, Gabor Lugosi, *Theory of classification: a survey of some recent advances*, ESAIM: PS 9 (2005), 323–375.

[5] A. Cohen, W. Dahmen, I. Daubechies and R. DeVore, *Tree-structured approximation and optimal encoding*, Appl. Comp. Harm. Anal. **11** (2001), 192-226.

[6] C. Scott and R. Nowak, *Minimax-optimal classification with dyadic decision trees*, IEEE Transactions on Information Theory **52** (2006), 1335–1353.

Learning sets with separating kernels

Lorenzo Rosasco
(joint work with Ernesto De Vito, Alessandro Toigo)

The setting we consider is described by a probability space (X, ρ) and a measurable reproducing kernel K on the set X [1]. The data are independent and identically distributed (i.i.d.) samples $x_1, \ldots, x_n$, each one drawn from X with probability ρ. The reproducing kernel K reflects some prior information on the problem and, as we discuss in the following, will also define the geometry of X. The goal is to use the sample points $x_1, \ldots, x_n$ to estimate the region where the probability measure ρ is concentrated.

To fix some ideas, the space X can be thought of as a high-dimensional Euclidean space and the distribution ρ as being concentrated on a region X_ρ, which is a smaller – and potentially lower dimensional – subset of X (e.g. a linear subspace or a manifold). In this example, the goal is to build from data an estimator X_n which is, with high probability, close to X_ρ with respect to a suitable metric.

We first note that a precise definition of X_ρ requires some care. If ρ is assumed to have a density with respect to some fixed reference measure (for example, the Lebesgue measure in the Euclidean space), then the region X_ρ can be easily defined to be the set of points where the density function is non-zero (or its closure). Nevertheless, this assumption would prevent considering the situation where the data are concentrated on a "small", possibly lower dimensional, subset of X. Note that, if the set X were endowed with a topological structure and ρ were defined on the corresponding Borel σ-algebra, it would be natural to define X_ρ as the support of the measure ρ, i.e. the smallest *closed* subset of X having measure one. However, since the set X is only assumed to be a measurable space, no a priori given topology is available. Here we also remark that the definition of X_ρ is not the only point where some further structure on X would be useful. Indeed, when defining a learning error, a notion of distance between the set X_ρ and its estimator X_n is also needed and hence some metric structure on X is required.

Now, the idea is to use the properties of the reproducing kernel K to induce a metric structure – and consequently a topology – on X. Indeed, under some mild technical assumptions on K, the function

$$d_k(x, y) = \sqrt{K(x, x) + K(y, y) - 2 \operatorname{Re} K(x, y)} \qquad \forall\, x, y \in X$$

defines a metric on X, thus making X a topological space. Then, it is natural to define X_ρ to be the support of ρ with respect to such metric topology. Note that the metric d_k also provides us with a notion of distance between closed sets, namely the corresponding Hausdorff distance d_H.

The problem we are interested in can now be restated in the following way: we want to learn from data an estimator X_n of X_ρ, such that $\lim_{n\to\infty} d_H(X_n, X_\rho) = 0$ almost surely. While X_ρ is now well defined, it is not clear how to build an estimator from data. A main result in the paper provides a new analytic characterization

of X_ρ, which immediately suggests a new computational solution for the corresponding learning problem. To derive and state this result, we introduce a new notion of reproducing kernels, called separating kernels, that, roughly speaking, captures the sense in which the reproducing kernel and the probability distribution need to be related. We say that a reproducing kernel Hilbert space $\mathcal{H}$ (or equivalently its kernel) *separates* a subset $C \subset X$, if, for any $x \notin C$, there exists $f \in \mathcal{H}$ such that

$$f(x) \neq 0 \quad \text{and} \quad f(y) = 0 \quad \forall y \in C.$$

If K separates all possible closed subsets in X, we say that it is *completely separating*.

Our main theorem states that, if either K is completely separating, or at least separates X_ρ, then X_ρ is the level set of a suitable distribution dependent continuous function F_ρ. More precisely, let $\mathcal{H}$ be the reproducing kernel Hilbert space associated to K [1], $T : \mathcal{H} \to \mathcal{H}$ the integral operator with kernel K, and denote by $T^\dagger$ its pseudo-inverse. If we consider the function F_ρ on X, defined by $F_\rho(x) = \langle T^\dagger T K_x, K_x \rangle \qquad \forall x \in X$, and K separates X_ρ, then we prove that $X_\rho = \{x \in X \mid F_\rho(x) = 1\}$, (where for simplicity we are assuming $K(x, x) = 1$ for all $x \in X$).

The above result is crucial since the integral operator T can be approximated with high probability from data (see [3] and references therein). However, since the definition of F_ρ involves the pseudo-inverse of T, the support estimation problem is ill-posed and regularization techniques are needed to ensure stability. With this in mind, we propose and study a family of spectral regularization techniques which are classical in inverse problems and have been considered in supervised learning in [2]. We define an estimator by

$$X_n = \{x \in X \mid F_n(x) \geq 1 - \tau_n\},$$

where $F_n(x) = (1/n)\mathbf{K}_{\mathbf{x}}^* g_{\lambda_n}(\mathbf{K}_n/n)\mathbf{K}_{\mathbf{x}}$, with $(\mathbf{K}_n)_{i,j} = K(x_i, x_j)$, $\mathbf{K}_{\mathbf{x}}$ is the column vector whose i-th entry is $K(x_i, x)$, and $\mathbf{K}_{\mathbf{x}}^*$ is its conjugate transpose. Here $g_{\lambda_n}(\mathbf{K}_n/n)$ is a matrix defined via spectral calculus by a spectral filter function g_{λ_n} that suppresses the contribution of the eigenvalues smaller than λ_n. Examples of spectral filters include Tikhonov regularization and truncated singular values decomposition, to name only a few. The error analysis for this class of methods can be derived in a unified framework and is done both in terms of asymptotic convergence, and stability to random sampling by means of finite sample bounds. Indeed, we prove that, if X is compact[1], then

$$\lim_{n\to\infty} \sup_{x\in X} |F_\rho(x) - F_n(x)| = 0 \qquad \text{almost surely,}$$

provided that $\lim_{n\to\infty} \lambda_n = 0$ and $\sup_{n\geq 1}(L_{\lambda_n} \log n)/\sqrt{n} < +\infty$, where L_{λ_n} is the Lipshitz constant of the function $r_{\lambda_n}(\sigma) = \sigma g_{\lambda_n}(\sigma)$. Moreover

$$\lim_{n\to\infty} d_H(X_n, X_\rho) = 0 \qquad \text{almost surely,}$$

[1]If X is not compact, these results hold replacing X with the intersection $X \cap C$ for any compact subset C.

provided that $\lim_{n\to\infty} \tau_n = 0$ and

$$\limsup_{n\to\infty} \frac{\sup_{x\in X} |F_n(x) - F_\rho(x)|}{\tau_n} \le 1 \qquad \text{almost surely.}$$

Note that, if X_ρ is separated by K, then the convergence of F_n to F_ρ can be proved without further assumptions on the problem. On the contrary, in order to have convergence of X_n to X_ρ we need to choose a sequence τ_n satisfying the condition above, and this requires knowledge of the convergence rate of F_n to F_ρ. The latter is a property of the couple (ρ, K), not only of K. If the couple is such that $\sup_{x\in X} \|T^{-s} K_x\| < \infty$, with $0 < s \le 1$, and the eigenvalues of the (compact and positive) operator T satisfy $\sigma_j \sim j^{-1/b}$ for some $0 < b \le 1$, then we prove that, for $n \ge 1$ and $\delta > 0$, we have

$$\sup_{x\in X} |F_n(x) - F_\rho(x)| \le C_{s,b,\delta} \left(\frac{1}{n}\right)^{\frac{s}{2s+b+1}}$$

with probability at least $1 - 2e^{-\delta}$, for $\lambda_n = n^{-1/(2s+b+1)}$ and a suitable constant $C_{s,b,\delta}$ which does not depend on n.

Finally, we remark that our construction relies on the assumption that the kernel K separates the support X_ρ. The question then arises whether there exist kernels that can separate a large number of, and perhaps all, closed subsets, namely kernels that are *completely separating*. The answer is affirmative, and for translation invariant kernels on $\mathbb{R}^d$, indeed a sufficient condition for a kernel to be completely separating can be given in terms of its Fourier transform. As a consequence, the Abel kernel $K(x, y) = e^{-\|x-y\|/\sigma}$ on the Euclidean space $X = \mathbb{R}^d$ is completely separating. Interestingly, the Gaussian kernel $K(x, y) = e^{-\|x-y\|^2/\sigma^2}$, which is very popular in machine learning, is not.

References

[1] N. Aronszajn *Theory of reproducing kernels*, T. Am. Math. Soc. **68** (1950), 337–404.

[2] F. Bauer, S. Pereverzev, and L. Rosasco, *On regularization algorithms in learning theory*, J. Complexity **23** (2007), 52–72.

[3] L. Rosasco, M. Belkin, and E. De Vito, *On learning with integral operators*, J. Mach. Learn. Res. **11** (2010), 905–934.

Irregular sampling in subspaces of $L_2(\mathbb{R})$

Joachim Stöckler

(joint work with Karlheinz Gröchenig)

The sampling theorem of Whittaker and Shannon states that a band-limited function $f \in L_2(\mathbb{R})$ is uniquely determined by its function values $f|_{\mathbb{Z}}$, in terms of the cardinal series

$$f(t) = \sum_{k\in\mathbb{Z}} f(k) \frac{\sin \pi(t-k)}{\pi(t-k)}.$$

Moreover, the identity $\|f\|_2^2 = \sum_{k\in\mathbb{Z}} |f(k)|^2$ holds. We present new results on the reconstruction of functions from a reproducing kernel Hilbert space $V \subset L_2(\mathbb{R})$, which can be a shift-invariant space of spline functions, or the image set of a bounded linear projection, or the linear span of irregular shifts of a given function. The given data are obtained by sufficiently dense nonuniform sampling; more precisely, we give conditions on the set $X \subset \mathbb{R}$ of sampling sites, such that the norm equivalence $\|f\|_2^2 \sim \sum_{x\in X} |f(x)|^2$ is satisfied. In this case, X is called a *set of sampling* for V.

Typical geometric conditions on the set X require that the maximal gap

$$\max_{y\in\mathbb{R}} \min_{x\in X} |y - x|$$

should be small as compared to some underlying structure of the space V. For example, if the functions in V satisfy a Bernstein-type inequality, then the method of "norming sets" can be applied in order to give sufficient conditions for X being a set of sampling. Only for very particular cases of shift-invariant spline spaces V, Aldroubi and Gröchenig [1] gave almost sharp conditions for X being a set of sampling. We extend these results to subspaces $V \subset L_2(\mathbb{R})$ which are spanned by irregular shifts of totally positive functions of finite type. This class of functions appears in the work of Schoenberg [2].

References

[1] A. Aldroubi, K. Gröchenig, *Beurling-Landau-type theorems for non-uniform sampling in shift invariant spline spaces*, J. Fourier Anal. Appl. **6** (2000), 93–103.

[2] I. J. Schoenberg, *On totally positive functions, Laplace integrals and entire functions of the Laguerre-Pólya-Schur type*, Proc. Natl. Acad. Sci. USA **33** (1947), 11–17.

Linear-phase moments in wavelet analysis and approximation theory

Bin Han

Approximation order of a shift invariant space generated by a single generating function is linked to the Strang-Fix condition, while the order of linear-phase moments of the generating function controls how the polynomials are exactly reproduced by the integer shifts of the generating function. We say that a compactly supported function $\phi \in L_1(\mathbb{R}^d)$ has the linear-phase moments of order m with phase c_ϕ if

$$\widehat{\phi}(\xi) = e^{-ic_\phi\cdot\xi} + O(\|\xi\|^m), \qquad \xi \to 0.$$

By Π_{m-1} we denote the space of all d-variate polynomials of total degree no more than $m-1$. Then $p * \phi := \sum_{k\in\mathbb{Z}^d} p(k)\phi(\cdot - k) = p(\cdot - c)$ for all $p \in \Pi_{m-1}$ if and only if ϕ has the linear-phase moments of order m with phase c and ϕ satisfies the Strang-Fix condition:

$$\partial^\mu \widehat{\phi}(2\pi k) = 0, \qquad \forall |\mu| < m, k \in \mathbb{Z}^d\backslash\{0\}.$$

The notion of linear-phase moments appeared initially but implicitly in [5] and has been formally introduced in [3, 4]. It has been further discussed for orthogonal wavelets and tight framelets in [2]. It turns out that linear-phase moments also play an interesting role in wavelet analysis. In this talk we discuss three applications of linear-phase moments in wavelet analysis: (1) Subdivision schemes with linear-phase moments which produce nearly shifted interpolatory subdivision schemes, (2) Role of linear-phase moments in the construction of symmetric tight framelets, (3) Linear-phase moments in the design of orthogonal filters used the Dual-Tree Complex Wavelet Transform (DT-CWT), which significantly outperforms the commonly used tensor product wavelets in signal and image processing.

Let M denote a $d \times d$ integer matrix and $a : \mathbb{Z}^d \to \mathbb{C}$ be a finitely supported sequence, called a filter. Define $\widehat{a}(\xi) := \sum_{k\in\mathbb{Z}^d} a(k)e^{-ik\cdot\xi}$. We say that a filter a has linear-phase moments of order m with phase $c_a \in \mathbb{R}^d$ if

$$\widehat{a}(\xi) = e^{-ic_a\cdot\xi} + O(\|\xi\|^m), \qquad \xi \to 0.$$

The subdivision operator $S_{a,M}$ is defined to be

$$[S_{a,M}v](m) := |\det(M)| \sum_{k\in\mathbb{Z}^d} v(k)a(m - Mk), \quad m \in \mathbb{Z}^d.$$

Let $a : \mathbb{Z}^d \to \mathbb{C}$, $c \in \mathbb{R}^d$, and m be an integer. Then it has been shown in [2, 5] that $S_{a,M}p = p(M^{-1}(\cdot - c))$ for all $p \in \Pi_{m-1}$ if and only if

(1) a has order m sum rules: $\widehat{a}(\xi + 2\pi\omega) = O(\|\xi\|^m)$ as $\xi \to 0$ for all $0 \neq \omega \in \Omega_M := ((M^T)^{-1}\mathbb{Z}^d) \cap [0, 1)^d$.
(2) a has the linear-phase moments of order m with phase c:

$$\widehat{a}(\xi) = e^{-ic\cdot\xi} + O(\|\xi\|^m), \qquad \xi \to 0.$$

Define $a^\star(k) := \overline{a(-k)}$ for all $k \in \mathbb{Z}^d$ and the convolution $[a^\star * a](n) := \sum_{k\in\mathbb{Z}^d} a^\star(n - k)a(k)$ for $n \in \mathbb{Z}^d$. The order of linear-phase moments of $a^\star * a$ is directly connected to the vanishing moments and therefore the frame approximation order of a tight framelet filter bank. The role of the linear-phase moments for symmetric tight framelet filters follows from the following fact. Suppose that a filter a has symmetry: $a(c_a - k) = \overline{a(k)}$ for $k \in \mathbb{Z}^d$ with $c_a \in \mathbb{Z}^d$. Then the correlation filter $a^\star * a$ has linear-phase moments of order m if and only if the filter a has the linear-phase moments of order m.

At the end of the talk, we shall also provide two other approaches for achieving directional representation. In the first approach, we show that using tensor product and using univariate complex-valued tight framelets we can obtain a 2D tight framelet with 4 directions: 0 degree, 45 degree, 90 degree, and 135 degree. For the second approach, we provide a tight framelet filter bank having an associated filter bank and having increasing number of directional obeying the hyperbolic rule. See [1] for more details. Such new directional tight framelets in 2D are expected to have applications in image processing.

References

[1] B. Han, *Nonhomogeneous wavelet systems in high dimensions*, Applied and Computational Harmonic Analysis **32** (2012), 169–196.
[2] B. Han, *Symmetric orthogonal filters and wavelets with linear-phase moments*, Journal of Computational and Applied Mathematics **236** (2011), 482–503.
[3] B. Han, *Symmetric orthonormal complex wavelets with masks of arbitrarily high linear-phase moments and sum rules*, Advances in Computational Mathematics **32** (2010), 209–237.
[4] B. Han, *Matrix extension with symmetry and applications to symmetric orthonormal complex M-wavelets*, Journal of Fourier Analysis and its Applications **15** (2009), 684–705.
[5] B. Han, *Vector cascade algorithms and refinable function vectors in Sobolev spaces*, Journal of Approximation Theory **124** (2003), 44–88.

Sampling scattered data with Bernstein polynomials: stochastic and deterministic error estimates

Zongmin Wu
(joint work with Xingping Sun, Limin Ma)

As an example to study the approximation of density functions, we first discuss different kinds of uniformly distributed points, including the classical uniformly distributed points, the quasi uniformly distributed points, the discrepancy, and the random uniformly distributions. The relations among them are studied. Especially, we have the following result.

Proposition. The following two statements are equivalent.

(1) The discrepancy D_n satisfies

$$D_n = \mathcal{O}(\frac{1}{n^\beta}).$$

(2) The following inequality

$$|x_j^n - \frac{j}{n}| = \mathcal{O}(\frac{1}{n^\beta})$$

holds true for each j.

Therefore, the random uniformly distributed points cannot be dominated by any discrepancy, since the inequality in the above item (2) will not be valid.

Viewing the classical Bernstein polynomials as sampling operators, we study a generalization by allowing the sampling operation to take place at scattered sites. We utilize both stochastic and deterministic approaches. On the stochastic side, we consider the sampling sites as random variables that obey some naturally derived probabilistic distributions, and obtain Chebyshev type estimates. On the deterministic side, we incorporate the theory of uniform distribution of point sets

(within the framework of Weyl's criterion) and the discrepancy method. We establish convergence results and error estimates under practical assumptions on the distribution of the sampling sites.

Theorem 1. Let $\{B_j^n(x) = \frac{n!}{j!(n-j)!}x^j(1-x)^{n-j}\}_{j=0}^n$ be the Bernstein basis, and ω be the modulus of continuity of a continuous function $f \in C([0,1])$. If $\frac{5}{2}\omega(\frac{1}{\sqrt{n}}) < \epsilon/2$, then the generalized quasi-interpolation that the data sampled on uniform random numbers

$$B_n^{**}f(x) = \sum_{j=0}^{n} f(x_j^n)B_j^n(x)$$

converges to f in probability on $[0,1]$, and the error can be bounded by

$$P\{\|B_n^{**}f(x) - f(x)\|_\infty > \epsilon\} \le 16/\epsilon^4 n^3 \omega(1/\sqrt{n})^4 \le 16/\epsilon^4 n.$$

The Bernstein basis B_j^n can be replaced by $\sqrt{n}W(\sqrt{n}(x-x_j^n))$, with any function satisfying $\sqrt{n}W(\sqrt{n}(x-t)) \to \delta(x-t)$. e.g.

Theorem 2. If $\{x_j^n\}$ are quasi uniform distributed points with the discrepancy $|x_j^n - j/n| < 1/n^\beta$, then

$$\begin{aligned}|\sum f(x_j^n)\sqrt{n}W(\sqrt{n}x - j/\sqrt{n}) - f(x)| \\ = \mathcal{O}(1/\sqrt{n}) + \omega(1/\sqrt{n}) + \omega(1/n^\beta)\end{aligned}$$

Theorem 3. Assume $\frac{5}{2}\omega(\frac{1}{\sqrt{n}}) < \epsilon/2$. Then the generalized quasi-interpolation that the data sampled on uniform random numbers

$$M_n^{**}f(x) = \sum_{j=0}^{n} f(x_j^n) \sum f(x_j^n)\sqrt{n}W(\sqrt{n}x - j/\sqrt{n})$$

converges to f in probability on $[0,1]$, and the error can be bounded by

$$P\{|M_n^{**}f(x) - f(x)\|_\infty > \epsilon\} \le 16/\epsilon^4 n^3 \omega(1/\sqrt{n})^4 \le 16/\epsilon^4 n.$$

We can generalize the above results to other domains with non classical standard uniformly distributed points, such as the surface of a ball and the disc. Similar results on error analysis are valid.

Furthermore we can discuss the approximation of density functions and probability distribution functions using statistical distances. The statistical distance is defined to minimize the energy cost to move the earth of one density function to another. Since the statistical distance for the univariate problem is equivalent to the L_1-norm of the inverse function of the probability distribution function, a

Bernstein scheme to approximate the inverse function of the probability distribution function is given in [1]. Error estimates for the statistical distance of such kind of approximation are given, which yield

Theorem 4. For random points $\{x_j^n\}$ drawn according to the unknown probability distribution function F, let $x = G(y)$ be the inverse function of F, then the Bernstein like scheme $G^*(y) = \sum x_j^n B_j^n(y)$ converges in probability to G. This concludes that it converges in probability with respect to the statistical distance.

References

[1] Zongmin Wu, Xingping Sun, Limin Ma, *Sampling scattered data with Bernstein polynomials: stochastic and deterministic error estimates*, Advances in Computational Mathematics, DOI 10.1007/s10444-011-9233-0 (2011).

Bernstein-Durrmeyer operators with arbitrary weight functions

Elena E. Berdysheva

Let ρ be a non-negative bounded (regular) Borel measure on the simplex

$$\mathbb{S}^d := \{x = (x_1, \dots, x_d) \in \mathbb{R}^d \ : \ 0 \le x_1, \dots, x_d \le 1, \ x_1 + \cdots + x_d \le 1\}.$$

We assume that $\operatorname{supp}\rho \setminus \partial\mathbb{S}^d \neq \emptyset$. The Bernstein basis polynomials of degree $n \in \mathbb{N}$ are defined by

$$B_\alpha(x) := \frac{n!}{\alpha_0!\alpha_1!\cdots\alpha_d!}\,(1 - x_1 - \cdots - x_d)^{\alpha_0}\, x_1^{\alpha_1} \cdots x_d^{\alpha_d},$$

where $\alpha_0, \alpha_1, \dots, \alpha_d \in \mathbb{N} \cup \{0\}$ and $\alpha_0 + \alpha_1 + \cdots + \alpha_d = n$. We introduce the Bernstein-Durrmeyer operator with weight ρ

$$\mathbf{M}_{n,\rho} f := \sum_{\alpha_0+\cdots+\alpha_d=n} \frac{\int_{\mathbb{S}^d} f\, B_\alpha \, d\rho}{\int_{\mathbb{S}^d} B_\alpha \, d\rho}\, B_\alpha \tag{1}$$

for $f \in L^q(\mathbb{S}^d, \rho)$, $1 \le q < \infty$, or $f \in C(\mathbb{S}^d)$. The operator $\mathbf{M}_{n,\rho}$ is linear and positive, and it reproduces constant functions. It generalizes the well-known Bernstein-Durrmeyer operators with Jacobi weights. A motivation for this generalization comes from learning theory: K. Jetter and D.-X. Zhou [3] have applied the univariate Bernstein-Durrmeyer operators of type (1) to bias-variance estimates for support vector machine classifiers. To our knowledge, [1] is the first paper where operators (1) in full generality were systematically investigated.

In the talk, we concentrate on discussing convergence of the operator $\mathbf{M}_{n,\rho}$. As a first step in studying convergence, we consider uniform convergence. We give necessary and sufficient conditions that guarantee uniform convergence on $\mathbb{S}^d$ for each function continuous on $\mathbb{S}^d$. Recall that a measure ρ on $\mathbb{S}^d$ is called strictly positive if $\rho(A \cap \mathbb{S}^d) > 0$ for every open set $A \subset \mathbb{R}^d$ such that $A \cap \mathbb{S}^d \neq \emptyset$. This is equivalent to the fact that $\operatorname{supp}\rho = \mathbb{S}^d$.

Theorem 1. [2] *Let ρ be a non-negative bounded Borel measure on $\mathbb{S}^d$ such that* $\operatorname{supp}\rho \setminus \partial\mathbb{S}^d \neq \emptyset$. *Then*

$$\lim_{n\to\infty} \|f - \mathbf{M}_{n,\rho} f\|_{C(\mathbb{S}^d)} = 0 \quad \textit{for every} \quad f \in C(\mathbb{S}^d)$$

if and only if ρ is strictly positive on $\mathbb{S}^d$.

A further natural question is about rates of convergence. The method used in [2] for arbitrary measures does not lead to estimates for rates of convergence. In [1], Jetter and the author obtained estimates for rates of convergence for the so-called Jacobi-like measures, i.e., for absolutely continuous measures ρ of the form

$$d\rho(x) = w(x)\,dx$$

such that

$$a\,(1-x_1-\cdots-x_d)^{\nu_0}\,x_1^{\nu_1}\cdots x_d^{\nu_d} \le w(x) \le A\,(1-x_1-\cdots-x_d)^{\mu_0}\,x_1^{\mu_1}\cdots x_d^{\mu_d}, \quad x \in \mathbb{S}^d,$$

with some $\nu = (\nu_0, \nu_1, \ldots, \nu_d)$, $\mu = (\mu_0, \mu_1, \ldots, \mu_d)$, where $\nu_i, \mu_i > -1$, $i = 0, 1, \ldots, d$, and $0 < a, A < \infty$. Obviously, Jacobi-like measures are strictly positive. The following statement follows from results of [1].

Theorem 2. *Let ρ be a Jacobi-like measure with $|\nu| - |\mu| < 1$. Let $f \in C(\mathbb{S}^d)$. Then*

$$\|f - \mathbf{M}_{n,\rho} f\|_{C(\mathbb{S}^d)} \le C\,\omega\left(f, n^{-\frac{1-(|\nu|-|\mu|)}{4}}\right),$$

where $\omega(f,\delta) = \sup\{|f(x) - f(t)| : \|t - x\|_2 < \delta\}$ denote the modulus of continuity of f.

As a next question, we study convergence of $\mathbf{M}_{n,\rho}$ in case when ρ is not strictly positive. We consider pointwise convergence on the support of the measure.

Theorem 3. *Let $x \in (\operatorname{supp}\rho)^\circ$. Let f be bounded on* $\operatorname{supp}\rho$ *and continuous at x. Then*

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

Recently, Bing-Zheng Li proved a statement about convergence of $\mathbf{M}_{n,\rho}$ in the spaces $L^q(\mathbb{S}^d, \rho)$.

Theorem 4. [4] *Let ρ be a non-negative bounded Borel measure on $\mathbb{S}^d$ such that* $\operatorname{supp}\rho \setminus (\partial\mathbb{S}^d) \neq \emptyset$. *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 $f \in L^q(\mathbb{S}^d, \rho)$. Moreover, for the functions $\varphi_i = x_i$, $i = 1, \ldots, d$, we have

$$\|\mathbf{M}_{n,\rho}(|\varphi_i(\cdot) - \varphi_i(x)|)\|_{L^q(\mathbb{S}^d,\rho)} \le \frac{1}{\sqrt{n}}, \qquad q = 1, 2.$$

Using a modification of her method, we can show that

$$\|\mathbf{M}_{n,\rho}(|\varphi_i(\cdot) - \varphi_i(x)|)\|_{L^q(\mathbb{S}^d,\rho)} \le \frac{C}{\sqrt{n}}, \qquad 1 \le q < \infty.$$

It follows that for $f \in L^q(\mathbb{S}^d, \rho)$ we have $\|f - \mathbf{M}_{n,\rho} f\|_{L^q(\mathbb{S}^d,\rho)} \le 2\, K_p\left(f, \frac{C}{\sqrt{n}}\right)$, where $K_p(f,t) = \inf\{\|f-g\|_{L^q(\mathbb{S}^d,\rho)} + t \max_{i=1,\dots,d} \|\partial_i g\|_C : g \in C^1(\mathbb{S}^d)\}$.

Part of the results was obtained jointly with Kurt Jetter.

References

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

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

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

[4] B.-Z. Li, *Approximation by multivariate Bernstein-Durrmeyer operators and learning rates of least-square regularized regression with multivariate polynomial kernels*, preprint (2012).

Error analysis and sparsity of some learning algorithms

Ding-Xuan Zhou

Sparsity is a classical topic in support vector machines of learning theory and is related to other research areas such as LASSO in statistics and compressed sensing. In this talk we discuss error analysis and sparsity for three kernel-based learning algorithms in a regression setting: support vector regression, coefficient-based regularization with ℓ^1-penalty, and kernel projection machines with ℓ^q-penalty.

Let X be a compact metric space (input space), $Y = \mathbb{R}$ (output space) and ρ be a probability measure on $Z = X \times Y$. Take a random sample $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m$ 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 some learning algorithms 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 fundamental functions $\{K_x = K(\cdot, x) : x \in X\}$.

The first learning algorithm we discuss is the support vector regression [1] defined by

$$f_{\mathbf{z}}^{\mathrm{SVR}} = \arg\min_{f \in \mathcal{H}_K} \left\{ \frac{1}{m} \sum_{i=1}^m \psi^\epsilon(f(x_i) - y_i) + \lambda \|f\|_K^2 \right\}, \tag{1}$$

where $\psi^\epsilon : \mathbb{R} \to \mathbb{R}_+$ is the ϵ-insensitive loss defined for $\epsilon \ge 0$ by

$$\psi^\epsilon(u) = \max\{|u| - \epsilon, 0\} = \begin{cases} |u| - \epsilon, & \text{if } |u| \ge \epsilon, \\ 0, & \text{otherwise.} \end{cases}$$

The original motivation for the insensitive parameter $\epsilon > 0$ is to balance the approximation ability and sparsity of the algorithm. The parameter changes with the sample size and usually $\epsilon = \epsilon(m) \to 0$ as the sample size m increases.

Algorithm (1) learns the median function $f_{\rho,\frac{1}{2}}$ on X which is defined by

$$\rho(\{y \in Y : y \le f_{\rho,\frac{1}{2}}(x)\}|x) \ge \frac{1}{2}, \qquad \rho(\{y \in Y : y \ge f_{\rho,\frac{1}{2}}(x)\}|x) \ge \frac{1}{2}.$$

The following learning rate is given in [2] for the output function $f_{\mathbf{z}}^{\mathrm{SVR}}$ after projecting the values onto the interval $[-M, M]$ under a noise condition given in [3]. Denote ρ_X as the marginal distribution ρ on X.

Theorem 1. Let $X \subset \mathbb{R}^n$ and $K \in C^\infty(X \times X)$. Assume $f_{\rho,\frac{1}{2}} \in \mathcal{H}_K$, $|y| \le M$ almost surely, and ρ has a median of p-average type 2 for some $p \in (0, \infty]$. Take $\lambda = m^{-\frac{p+1}{p+2}}$ and $\epsilon = m^{-\beta}$ with $\frac{p+1}{p+2} \le \beta \le \infty$. Let $0 < \eta < \frac{p+1}{2(p+2)}$. Then with $p^* = \frac{2p}{p+1} > 0$, with confidence $1 - \delta$, we have

$$\left\| \pi_M(f_{\mathbf{z}}^{\mathrm{SVR}}) - f_{\rho,\frac{1}{2}} \right\|_{L^{p^*}_{\rho_X}} \le \widetilde{C} \log \frac{3}{\delta} m^{\eta - \frac{p+1}{2(p+2)}},$$

where $\widetilde{C}$ is a constant independent of m or δ. Here the condition that ρ has a median of p-average type 2 means there exist $a_x \in (0, 2]$, $b_x > 0$ such that the function $\frac{1}{b_x a_x}$ lies in $L^p_{\rho_X}$ and for each $u \in [0, a_x]$, the $\rho(\cdot|x)$ measures of $(f_{\rho,\frac{1}{2}}(x) - u, f_{\rho,\frac{1}{2}}(x))$ and $(f_{\rho,\frac{1}{2}}(x), f_{\rho,\frac{1}{2}}(x) + u)$ are both at least $b_x u$.

The second algorithm we discuss is the coefficient-based regularization with ℓ^1-penalty defined as $f_{\mathbf{z},\lambda} = \sum_{k=1}^m \alpha^{\mathbf{z}}_{\lambda,k} K_{x_k}$, where $\alpha^{\mathbf{z}}_\lambda = (\alpha^{\mathbf{z}}_{\lambda,k})_{k=1}^m$ is given by

$$\alpha^{\mathbf{z}}_\lambda = \arg\min_{\alpha \in \mathbb{R}^m} \left\{ \frac{1}{m} \sum_{i=1}^m \left(\sum_{k=1}^m \alpha_k K_{x_k}(x_i) - y_i \right)^2 + \lambda \|\alpha\|_1 \right\}. \tag{2}$$

It is motivated by linear programming support vector machines and ridge regression [4]. Difficulty in error analysis caused by the sample dependence nature of the algorithm was estimated by local polynomial reproduction techniques developed in the literature of scattered data interpolation. The following learning rate can be found in [5]. Define an integral operator L_K on $L^2_{\rho_X}$ by $L_K f = \int_X K_u f(u) d\rho_X$.

Theorem 2 Assume that $X \subset \mathbb{R}^n$ has piecewise smooth boundary and satisfies an interior cone condition. Suppose ρ_X satisfies condition L_τ:

$$\rho_X\left(B(x, r)\right) \ge C_\tau r^\tau \qquad \forall x \in X, 0 < r \le 1$$

for some $\tau > 0$ and $C_\tau > 0$, $K \in C^\infty(X \times X)$ and f_ρ lies in the range of L^2_K. Let $0 < \epsilon < \frac{1}{2}$ and $\lambda = m^{\epsilon - \frac{1}{2}}$. Then for any $0 < \delta < 1$, with confidence $1 - \delta$, we have

$$\|f_{\mathbf{z},\lambda} - f_\rho\|^2_{L^2_{\rho_X}} \le \widetilde{C}_\epsilon \left\{ \log(m+1) + \log \frac{2}{\epsilon\delta} \right\}^{\frac{8}{\epsilon^2} + \frac{8n}{\tau\epsilon^2}} m^{\epsilon - \frac{1}{2}},$$

where $\widetilde{C}_\epsilon$ is a constant independent of m or δ.

The last algorithm we discuss is kernel projection machines with ℓ^q-penalty $(0 < q \le 1)$. Here we regard L_K as an integral operator on $\mathcal{H}_K$ with normalized eigenpairs $\{(\lambda_i, \phi_i)\}$. Its empirical version $L^{\mathbf{x}}_K : \mathcal{H}_K \to \mathcal{H}_K$ is defined by

$$L^{\mathbf{x}}_K f = \frac{1}{m} \sum_{i=1}^m f(x_i) K_{x_i} = \frac{1}{m} \sum_{i=1}^m \langle f, K_{x_i} \rangle_K K_{x_i}.$$

Denote its normalized eigenpairs as $\{(\lambda^{\mathbf{x}}_i, \phi^{\mathbf{x}}_i)\}$. Then the kernel projection machine with ℓ^q-penalty for regression [6] produces the output function $f^{\mathbf{z}} = \sum_{i=1}^\infty c^{\mathbf{z}}_i \phi^{\mathbf{x}}_i$

with $c^{\mathbf{z}} = (c_i^{\mathbf{z}})_{i=1}^{\infty}$ given by

$$c^{\mathbf{z}} = \arg\min_{c\in\ell^2} \left\{ \frac{1}{m}\sum_{i=1}^{m} \left(\sum_{j=1}^{\infty} c_j \phi_j^{\mathbf{x}}(x_i) - y_i \right)^2 + \gamma \|c\|_q^q \right\}, \qquad \gamma > 0. \tag{3}$$

The optimization problem (3) can be reduced to minimization of univariate functions. Denote $a_q = (1-q)^{1-q}\left(2/(2-q)\right)^{2-q}$ and $S_i^{\mathbf{z}} = \frac{1}{m\lambda_i^{\mathbf{x}}}\sum_{j=1}^{m} y_j \phi_i^{\mathbf{x}}(x_j)$, if $\lambda_i^{\mathbf{x}} > 0$, and $S_i^{\mathbf{z}} = 0$ otherwise.

Theorem 3 (a) $c_i^{\mathbf{z}} = \arg\min_{c\in\mathbb{R}} \left\{ \lambda_i^{\mathbf{x}}(c - S_i^{\mathbf{z}})^2 + \gamma |c|^q \right\}, \quad \forall i.$

(b) If $a_q \lambda_i^{\mathbf{x}} |S_i^{\mathbf{z}}|^{2-q} < \gamma$, then $c_i^{\mathbf{z}} = 0$.

If $a_q \lambda_i^{\mathbf{x}} |S_i^{\mathbf{z}}|^{2-q} = \gamma$, then $c_i^{\mathbf{z}} = 0$ or $\frac{2-2q}{2-q} S_i^{\mathbf{z}}$.

If $a_q \lambda_i^{\mathbf{x}} |S_i^{\mathbf{z}}|^{2-q} > \gamma$, then $c_i^{\mathbf{z}}$ is uniquely defined, has the same sign as $S_i^{\mathbf{z}}$ and satisfies $|S_i^{\mathbf{z}}| - (\gamma/\lambda_i^{\mathbf{x}})^{1/(2-q)} < |c_i^{\mathbf{z}}| < |S_i^{\mathbf{z}}|$. In the case $q = 1$, we have

$$c_i^{\mathbf{z}} = S_i^{\mathbf{z}} - \operatorname{sgn}(S_i^{\mathbf{z}}) \frac{\gamma}{2\lambda_i^{\mathbf{x}}}.$$

(c) $c_i^{\mathbf{z}} = 0$ if $\lambda_i^{\mathbf{x}} = 0$.

The following analysis [6] shows that sparsity of the algorithm improves while its learning ability is weakened as q decrease to 0.

Theorem 4 Assume $D_1 i^{-\alpha_1} \le \lambda_i \le D_2 i^{-\alpha_2}$ for every i, where $D_1, D_2 > 0$ and $\alpha_1 \ge \alpha_2 > \frac{2-q}{2(q(1+r)-1)}$. Let $0 < \delta < 1$, $\xi = \frac{q}{2\alpha_1 + 2\alpha_2(q(1+r)-1)} < 1$, and $\gamma = C_1 \left((\lambda_{\lceil m^\xi \rceil})^{q(r+1)} + \left(\log \frac{4}{\delta}\right)^{q(1+r)} m^{-\frac{q}{2}} \right)$. We have with confidence $1-\delta$ that

$$c_i^{\mathbf{z}} = 0, \qquad \forall m^\xi + 1 \le i \le m \tag{4}$$

and

$$\|f^{\mathbf{z}} - f_\rho\|_K \le C^* \left(\log \frac{4}{\delta} \right)^{1+r} m^{-\theta}, \tag{5}$$

where C_1, C^* are constants and

$$\theta = \frac{q\left(2\alpha_2(q(1+r)-1) - (2-q)\right)}{2(2-q)\left(2\alpha_1 + 2\alpha_2(q(1+r)-1)\right)} > 0. \tag{6}$$

References

[1] V. Vapnik, *Statistical Learning Theory*, Wiley, New York (1998).

[2] D. H. Xiang, T. Hu, and D. X. Zhou, *Approximation analysis of learning algorithms for support vector regression and quantile regression*, J. Appl. Math. **2012** (2012), Article ID 902139, 17 pages.

[3] I. Steinwart and A. Christman, *How SVMs can estimate quantile and the median*, Advances in Neural Information Processing Systems **20** (2008), 305-312, MIT Press, Cambridge, MA.

[4] Q. Wu and D. X. Zhou, *Learning with sample dependent hypothesis spaces*, Computers and Mathematics with Applications **56** (2008), 2896-2907.

[5] L. Shi, Y. L. Feng, and D. X. Zhou, *Concentration estimates for learning with ℓ^1-regularizer and data dependent hypothesis spaces*, Appl. Comput. Harmonic Anal. **31** (2011), 286–302.

[6] X. Guo and D. X. Zhou, *An empirical feature-based learning algorithm producing sparse approximations*, Appl. Comput. Harmonic Anal. **32** (2012), 389-400.

Stochastic aspects of nonlinear refinement algorithms

Oliver Ebner

Linear subdivision schemes of the form

$$Sx_i = \sum_{j\in\mathbb{Z}^s} a_{i-2j}x_j, \tag{1}$$

with $(a_i)_{i\in\mathbb{Z}^s}$ a finitely supported mask of nonnegative coefficients, are well-studied means to iteratively construct continuous functions from discrete data samples from a linear space. Loosely speaking, $S^n x$ gives a processed version of the input data $x \in \ell^\infty(\mathbb{R})$, which, interpreted as a function on $2^{-n}\mathbb{Z}^s$, should uniformly approximate a continuous function $S^\infty x : \mathbb{R}^s \to \mathbb{R}$. In the linear case, convergence and smoothness properties of subdivision schemes are well-understood. However, amounting to a deluge of data structurally confined to manifolds or, more generally, metric spaces, there has been a recent interest in the generalization of these kinds of refinement algorithms to the nonlinear setting. As uniform convergence of a subdivision scheme implies that $\sum_{j\in\mathbb{Z}^s} a_{i-2j} = 1$ for $i \in \mathbb{Z}^s$, a sensible generalization of (1) to metric spaces is given by

$$Sx_i = \operatorname{argmin}\Big(\sum_{j\in\mathbb{Z}^s} a_{i-2j} d^2(x_j, \cdot)\Big). \tag{2}$$

This type of refinement algorithm is referred to as **barycentric subdivision scheme**. It can be shown that barycentric schemes are well-defined on any simply connected, complete Alexandrov space of nonpositive curvature. On this class of metric spaces, referred to as **Hadamard spaces**, the subdivision rule (2) may be interpreted stochastically as follows. Note that the **subdivision matrix** $(a_{i-2j})_{i,j\in\mathbb{Z}^s}$ is row-stochastic and thus gives rise to a Markov chain X_n with transition probabilities $\mathbb{P}(X_{n+1} = j \mid X_n = i) = a_{i-2j}$. As a consequence of a nonlinear Markov property, the subdivision semigroup coincides with the Markov semigroup associated to X_n. More precisely, it holds that

$$S^n x \circ X_0 = E(x(X_n)|||(\mathfrak{F}_k)_{k\geq 0}),$$

where $E(\,\cdot\,|||(\mathfrak{F}_k)_{k\geq 0})$ denotes the **filtered conditional expectation** introduced by K.-T. Sturm. This observation, together with a nonlinear version of Jensen's inequality, paves the way for the a 'linear equivalence'-type theorem, stating that a barycentric refinement scheme converges on arbitrary Hadamard spaces **if and only if** it converges for real-valued input data, see [2].

References

[1] O. Ebner, *Convergence of refinement schemes on metric spaces*, Proc. Amer. Math. Soc. (2012), to appear.

[2] O. Ebner, *Stochastic aspects of refinement schemes on metric spaces*, technical report (2012), TU Graz.

[3] K.-T. Sturm, *Nonlinear martingale theory for processes with values in metric spaces of nonpositive curvature*, Ann. Prob. **30/3** (2002), 1195-1222.

Non-negative subdivision and Markov chains

Kurt Jetter
(joint work with Xianjun Li)

Recent work by X. L. Zhou, see [7] and the references there, has settled a long-standing question of characterizing uniform convergence of non-negative, univariate subdivision schemes with finitely supported masks. In a proper setting, going back to the seminal paper [5] of Micchelli and Prautzsch, convergence can be studied through properties of a related non-homogeneous Markov chain. We reprove and extend the existent convergence results for non-negative subdivision, in the multivariate version, by using the Anthonisse-Tijms result on convergence of such Markov chains.

Their analysis is based on the notion of SIA matrices, as introduced in [6], and on sign patterns of products of row stochastic matrices (or equivalently, properties of their directed graphs). Also, Hajnal's τ-coefficient of ergodicity, see [3] proves to be useful for studying the convergence of infinite products from a finite family of row stochastic matrices.

For scalar-valued subdivision, this approach to non-negative subdivsion can be found in the recent paper [4]. Beyond that, some of the ideas can be applied to matrix subdivision in a straightforward way.

Concerning the application of non-negative subdivision in spaces of Hadamard type, see O. Ebner's talk in this workshop, [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] O. Ebner, *Stochastic aspects of refinement schemes on metric spaces*, Dissertation, Technische Universität Graz, 2012.
[3] J. Hajnal, *Weak ergodicity in non-homogeneous Markov chains*, Proc. Cambridge Phil. Soc. **54** (1958), 233–246.
[4] K. Jetter and X. Li, *SIA matrices and non-negative subdivision*, Results in Mathematics (2012), to appear.
[5] C. A. Micchelli and H. Prautzsch, *Uniform refinement of curves*, Linear Algebra Appl. **114/115** (1989), 841–870.
[6] J. Wolfowitz, *Products of indecomposable, aperiodic, stochastic matrices*, Proc. Amer. Math. Soc. **14** (1963), 733–737.
[7] X.-L. Zhou, *Positivity of refinable functions defined by nonnegative masks*, Appl. Comput. Harmonic Analysis **27** (2009), 133–156.

Reporters: Kurt Jetter and Ding-Xuan Zhou

Participants

Prof. Dr. Misha Belkin
Department of Computer Science
and Engineering
Ohio State University
2015 Neil Ave.
Columbus , OH 43210-1277
USA

Prof. Dr. Elena Berdysheva
German University of Technology in Oman
(GUTech)
PO Box 1816, PC 130
Athaibah, Muscat
SULTANATE OF OMAN

Prof. Dr. Martin D. Buhmann
Lehrstuhl für Numerische Mathematik
Universität Gießen
Heinrich-Buff-Ring 44
35392 Giessen

Dr. Wolfgang zu Castell
Scientific Computing Research Unit
Helmholtz Zentrum München
Ingolstädter Landstrasse 1
85764 Neuherberg

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

Prof. Dr. Wolfgang Dahmen
Institut für Geometrie und
Praktische Mathematik
RWTH Aachen
Templergraben 55
52056 Aachen

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

Prof. Dr. Luc Devroye
School of Computer Science
McGill University
Montreal Quebec H3A 2K6
CANADA

Dr. Maik Döring
Institut für Angewandte Mathematik
und Statistik
Universität Hohenheim
Schloß, Westhof-Süd
70599 Stuttgart

Mona Eberts
Fachbereich Mathematik
Universität Stuttgart
Pfaffenwaldring 57
70569 Stuttgart

Prof. Dr. Oliver Ebner
Institut für Geometrie
TU Graz
Kopernikusgasse 24
A-8010 Graz

Prof. Dr. Laszlo Györfi
Department of Computer Science and
Information Theory
Budapest University of Techn.& Economics
Stoczek u. 2
H-1521 Budapest

Dr. Minh Ha Quang
Istituto Italiano di Tecnologia (IIT)
Via Moreago, 30
I-16163 Genova

Prof. Dr. Hakop Hakopian
Department of Informatics and
Applied Mathematics
Yerevan State University
A. Manoogian Str. 1
Yerevan 0025
ARMENIA

Prof. Dr. Bin Han
Department of Mathematical and
Statistical Sciences
University of Alberta
Edmonton, Alberta T6G 2G1
CANADA

Prof. Dr. Matthias Hein
FR 6.1 - Mathematik
Universität des Saarlandes
Postfach 15 11 50
66041 Saarbrücken

Prof. Dr. Kurt Jetter
Institut für Angewandte Mathematik
und Statistik
Universität Hohenheim
Schloß, Westhof-Süd
70599 Stuttgart

Prof. Dr. Jürgen Jost
Max-Planck-Institut für Mathematik
in den Naturwissenschaften
Inselstr. 22 - 26
04103 Leipzig

Prof. Dr. Gitta Kutyniok
Institut für Mathematik
Sekr. MA 4-1
Technische Universität Berlin
Straße des 17. Juni 136
10623 Berlin

Prof. Dr. Song Li
Department of Mathematics
Center for Mathematical Sciences
Zhejiang University
Hangzhou 310 027
P.R. CHINA

Prof. Dr. Lek-Heng Lim
Department of Statistics
The University of Chicago
5734 University Avenue
Chicago , IL 60637-1514
USA

Prof. Dr. Gabor Lugosi
Department of Economics
Pompeu Fabra University
Ramon Trias Fargas 25-27
E-08005 Barcelona

Dr. Ulrike von Luxburg
Fachbereich Informatik
Universität Hamburg
Vogt-Kölln-Str. 30
22527 Hamburg

Prof. Dr. Hrushikesh N. Mhaskar
Department of Mathematics
California Institute of Technology
Pasadena , CA 91125
USA

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

Prof. Dr. Sergei Pereverzyev
Johann Radon Institute
Austrian Academy of Sciences
Altenberger Straße 69
A-4040 Linz

Prof. Dr. Gerlind Plonka-Hoch
Institut für Numerische
und Angewandte Mathematik
Universität Göttingen
Lotzestr. 16-18
37083 Göttingen

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

Dr. Lorenzo Rosasco
MIT
Bldg. 46-5155
43 Vassar Street
Cambridge , MA 02139
USA

Prof. Dr. Robert Schaback
Institut für Numerische
und Angewandte Mathematik
Universität Göttingen
Lotzestr. 16-18
37083 Göttingen

Prof. Dr. Thomas Schick
Mathematisches Institut
Georg-August-Universität Göttingen
Bunsenstr. 3-5
37073 Göttingen

Prof. Dr. Nat Smale
Department of Mathematics
University of Utah
155 South 1400 East
Salt Lake City , UT 84112-0090
USA

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

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

Prof. Dr. Joachim Stöckler
Institut für Angewandte Mathematik
Technische Universität Dortmund
Vogelpothsweg 87
44227 Dortmund

Prof. Dr. Alexandre B. Tsybakov
Laboratoire de Probabilites
Universite Paris 6
4 place Jussieu
F-75252 Paris Cedex 05

Prof. Dr. Alessandro Verri
DISI
Universita di Genova
V. Dodecaneso 35
I-16146 Genova

Prof. Dr. Jean-Philippe Vert
Mines ParisTech
Centre for Computational Biology
35 rue Saint Honore
F-77305 Fontainebleau Cedex

Prof. Dr. Grace Wahba
Department of Statistics
University of Wisconsin
MSC, 1300 University Ave.
Madison WI 53706-1685
USA

Prof. Dr. Holger Wendland
Mathematical Institute
Oxford University
24-29 St. Giles
GB-Oxford OX1 3LB

Prof. Dr. Qiang Wu
Department of Mathematical Sciences
Middle Tennessee State University
Box 34
Murfreesboro , TN 37132-0001
USA

Prof. Dr. Zongmin Wu
School of Mathematical Sciences
Fudan University
220 Hadan Road
Shanghai 200 433
P.R.CHINA

Prof. Dr. Yuan Yao
Department of Mathematics
Peking University
Beijing 100 871
P.R.CHINA

Prof. Dr. Tong Zhang
Department of Statistics
Rutgers University
110 Frelinghuysen Road
Piscataway , NJ 08854-8019
USA

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

Dr. Xiaosheng Zhuang
Fachbereich Mathematik
Technische Universität Berlin
Straße des 17. Juni 136
10623 Berlin

Prof. Dr. Georg Zimmermann
Institut für Angewandte Mathematik
und Statistik
Universität Hohenheim
Schloß, Westhof-Süd
70599 Stuttgart