Mathematisches Forschungsinstitut Oberwolfach

Report No. 30/2008

Learning Theory and Approximation

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

June 29th – July 5th, 2008

Abstract. Mathematical analysis of learning algorithms consists of bias measured by various kinds of approximation errors and variance investigated by probability and statistical analysis. This workshop has dealt with new developments and achievements from the past ten years, such as sparsity and dimension reduction for huge dimensional data, kernel learning and approximation by integral operators, or non-linear approximation and learning by scaling.

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 (Berkeley) and Ding-Xuan Zhou (Hong Kong), was held June 29 – July 5, 2008. The meeting was attended by 22 participants from Europe, North America and Asia. It provided an excellent platform for a fruitful interaction of scientists from learning theory and approximation theory. Discussions with participants in the parallel workshop *Computational Algebraic Topology* were encouraged by the program to include plenary sessions for both workshops on subjects of common interest. Among these, the workshop addressed here has contributed six plenary lectures presented by Smale, Suykens, Tsybakov, Sauer, Temlyakov, and Schölkopf.

The scientific program started with Smale's lecture which was perfectly within the scope of both workshops. His idea was to apply various orders of boundary and co-boundary operators in Hodge theory to develop learning algorithms for pattern analysis on general probability spaces without Lebesgue structure, similar to the graph Laplacian. To this end, approximation theory is needed for the study of approximating integral operators associated with positive kernels by finite-rank

operators and of the space of harmonic functions on general spaces. A useful demonstration of this type of approximation of integral operators by finite-rank operators was given by Tarres in his talk dealing with online learning algorithms for regression.

Various central branches of learning theory have raised new approximation theory problems. For example, kernel-based learning has become an indispensable tool, and two plenary lectures have addressed this subject: Suykens surveyed various support vector machine type kernel-based learning algorithms and pointed to approximation theory problems about kernel canonical correlation analysis, independent component analysis and kernel PCA for sparsity. Schölkopf gave a general introductory survey on kernel methods and discussed several interesting approximation theory problems on data dependent kernels, how to measure variable dependence and covariance by kernel means, and learning surfaces from normals. However, kernel-based methods have been used also in approximation theory for some time, where explicit error estimates have been derived for the approximation of smooth classes of functions. Examples include the use of frame decompositions, see the contribution by Han on wavelet frames, or the use of a class of Bernstein-Durrmeyer positive linear operators associated with general probability measures and their applications to kernel learning algorithms, see the talks by Jetter and by Berdysheva.

A second central topic in learning theory is sparsity, which is an important property for dimension reduction, data representation and analysis, and information retrieval. In this workshop, some statisticians discussed sparsity for various purposes and raised the interesting problem of how to characterize functions with sparse representations: Tsybakov discussed how to study the sparsity in statistical learning theory by sparsity oracle inequalities. Wahba discussed LASSO algorithm and presented a separable approximation algorithm with projected Newton acceleration for variable selection and clustering. Pontil surveyed some multi-task learning with spectral regularization. Mukherjee talked about some predictive models inferring structure and dependencies. And Boucheron discussed model selection for Wilks phenomenon.

On the other hand, sparsity comes up in approximation theory with the development of algorithms for non-linear approximation. Binev presented a mathematical analysis of an adaptive sparse tree algorithm for regression problems in learning theory by such non-linear methods. Temlyakov talked about optimal error bounds of some universal estimators in learning theory by means of properties of N-term approximation. Sauer considered a multivariate polynomial interpolation problem to learn an algebraic variety containing a finite set of points for the purpose of dimension reduction. Zhou described some classification and regression learning schemes generated by Parzen windows and least squares regularization from a new viewpoint of scaling in the time and the frequency domain. And Rosasco introduced some learning algorithms associated with elastic net regularization by means of some classes of wavelets and libraries from approximation theory.

Two further talks have dealt with application of non-parametric regression: Kohler applied smooth splines to non-parametric regression estimators by Monte Carlo methods, with an interesting application to pricing of options. And Hein studied non-parametric regression between manifolds and asked for properties of thin-plate splines and higher order energy functionals including the so-called Eells energy.

The workshop has provided an excellent overview on actual subjects where learning theory and approximation theory meet, and where the two fields can benefit from each other. It has also raised some questions to be settled in the future. To address two of them: First, the usual estimates for variance involve essentially capacity of function classes used in learning processes which relies on properties of various functions spaces from approximation theory. However, smoothness of functions is usually measured in a sophisticated way which makes such results less applicable, in particular, if high-dimensional data are considered. This question is important to topics and learning algorithms on dimension reduction. Second, learning theory considers robustness of its methods, so far, only concerning statistical errors. However, from the standpoint of numerical analysis, also the condition of the used algorithms should be incorporated.

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

Hodge Decomposition and Learning Theory

Steve Smale
(joint work with Nat Smale)

Partial differential equations and Laplacians in Euclidean spaces together with the Lebesgue measure and its counterpart on manifolds have played a central role in understanding natural phenomena. In many areas, calculus is obstructed as in singular spaces, computer vision, learning theory, and quantum field theory. In vision it would be useful to do analysis on the space of images and an image is a function on a patch.

The point of view taken in this talk is to benefit from the Hodge theory to develop pattern analysis on general probability spaces without Lebesgue structure. This starts with a set X equipped with a distance d (which yields analysis like PDE and heat equations) as well as a probability measure ρ (measuring the distribution of objects like images in X).

Let $\ell \in \mathbb{Z}_+$. The space $L^2(X^{\ell+1}) = L^2_\rho(X^{\ell+1})$ consists of ℓ-forms. The Hodge operator or co-boundary $\delta : L^2(X^{\ell+1}) \to L^2(X^{\ell+2})$ is defined by

$$\delta f(x^0, \dots, x^{\ell+1}) = \sum_{i=0}^{\ell+1} (-1)^i f(x^0, \dots, \widehat{x^i}, \dots, x^{\ell+1}).$$

Its dual $\delta^* = \partial : L^2(X^{\ell+2}) \to L^2(X^{\ell+1})$ is called the boundary operator.

The Laplacian on ℓ-forms is defined to be the operator $\Delta : L^2(X^{\ell+1}) \to L^2(X^{\ell+1})$ given by $\Delta = \delta\partial + \partial\delta$. If we denote Harm to be the space of all harmonic functions in $L^2(X^{\ell+1})$ satisfying $\Delta f = 0$, then we have the following Hodge decomposition (L^2 theory) [1].

Theorem 1. $L^2(X^{\ell+1}) = \mathrm{Im}\partial + \mathrm{Im}\delta + \mathrm{Harm}$.

The Hodge operator δ can be generalized to a weighted setting with a symmetric and positive function K on $X \times X$. To see this, let $A_{\ell+1}$ be the weight function on $X^{\ell+1}$ given by $A_{\ell+1}(x^0, \dots, x^\ell) = \Pi_{i \neq j} \left(K(x^i, x^j)\right)^{1/2}$ for $\ell \geq 1$ while $A_1 \equiv 1$. Then the Hodge operator $\delta = \delta_K$ is from the weighted space $L^2_{\rho A_{\ell+1}}(X^{\ell+1})$ to the weighted space $L^2_{\rho A_{\ell+2}}(X^{\ell+2})$. Its dual $\delta^* = \partial : L^2_{\rho A_{\ell+2}}(X^{\ell+2}) \to L^2_{\rho A_{\ell+1}}(X^{\ell+1})$ is given by

$$\partial f(x^0, \dots, x^\ell) = \sum_{i=0}^{\ell+1} (-1)^i \int_X f(x^0, \dots, x^{i-1}, u, x^i, \dots, x^\ell) \Pi_{j=0}^\ell K(x^j, u) d\rho(u).$$

The Hodge operator and induced Laplacian can be used for learning theory. Consider the case $\ell = 0$ in the weighted setting with K being a Mercer kernel on

X. Then $A_1 \equiv 1$ and $A_2(x^0, x^1) = K(x^0, x^1)$. The Laplacian $\Delta = \partial\delta : L^2_\rho(X) \to L^2_\rho(X)$ on 0-forms takes the form

$$\Delta f(x) = 2D(x)f(x) - 2L_K f(x),$$

where $D(x) = \int_X K(x,u)d\rho(u)$ and L_K is the integral operator on $L^2_\rho(X)$ or the reproducing kernel Hilbert space $\mathcal{H}_K$ given by $L_K f(x) = \int_X K(x,u)f(u)d\rho(u)$.

The operator Δ can also be considered as one on $\mathcal{H}_K$. It can be discretized by a sample $\{x_i\}_{i=1}^m$ drawn from ρ. The function $D \in \mathcal{H}_K$ can be discretized as $\frac{1}{m}\sum_{i=1}^m K_{x_i}$ where $K_x = K(\cdot, x) \in \mathcal{H}_K$. The operator $L_K : \mathcal{H}_K \to \mathcal{H}_K$ can be approximated by a finite-rank one $\frac{1}{m}S_{\mathbf{x}}^T S_{\mathbf{x}}$ (induced by a sample operator $S_{\mathbf{x}}$ as in [2]) defined as $\frac{1}{m}S_{\mathbf{x}}^T S_{\mathbf{x}} f = \frac{1}{m}\sum_{i=1}^m \langle \cdot, K_{x_i}\rangle_K K_{x_i}$.

Theorem 2. *Assume* $\kappa := \sqrt{\sup_{x\in X} K(x,x)} < \infty$. *With confidence* $1-\delta$,

$$\left\| \frac{1}{m} S_{\mathbf{x}}^T S_{\mathbf{x}} - L_K \right\|_{\mathcal{H}_K \to \mathcal{H}_K} \le \frac{4\kappa^2 \log\big(2/\delta\big)}{\sqrt{m}}.$$

Consider another weighted setting (corresponding to adjacency matrix of a graph X). Let $\alpha > 0$ and a subset of $X^{\ell+1}$ given by $\mathcal{U}_\alpha^{\ell+1} = \{(x^0, \ldots, x^\ell) \in X^{\ell+1} : d(x^i, p) \le \alpha \text{ for some } p \in X, \text{ and all } i\}$ (it equals $X^{\ell+1}$ when α is large enough). The Hodge operator $\delta = \delta_\alpha$ can be regarded as one from $L^2_\rho(\mathcal{U}_\alpha^{\ell+1})$ to $L^2_\rho(\mathcal{U}_\alpha^{\ell+2})$. Its dual $\partial : L^2_\rho(\mathcal{U}_\alpha^{\ell+2}) \to L^2_\rho(\mathcal{U}_\alpha^{\ell+1})$ is given by

$$\partial f(x^0, \ldots, x^\ell) = \sum_{i=0}^{\ell+1} (-1)^i \int_{S_{x^0,\cdots,x^\ell}} f(x^0, \ldots, x^{i-1}, u, x^i, \ldots, x^\ell) d\rho(u).$$

Here $S_{x^0,\cdots,x^\ell}$ denotes the slice $\{t \in X : (x^0, \ldots, x^\ell, t) \in \mathcal{U}_\alpha^{\ell+2}\}$. In this setting we have the following Hodge decomposition [1] where the space Harm of harmonic functions is defined by the corresponding Laplacian.

Theorem 3. *For any* $\alpha > 0$ *and* $\ell \in \mathbb{Z}_+$, *we have*

$$L^2_\rho(\mathcal{U}_\alpha^{\ell+1}) = \mathrm{Im}\partial + \mathrm{Im}\delta + \mathrm{Harm}.$$

The space of harmonic functions and in general eigenfunctions of the above Laplacian would lead to some applications in pattern analysis [3] as the graph Laplacian [4] does.

References

[1] N. Smale and S. Smale, *A general Hodge theory*, in preparation.

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

[3] S. Smale and D. X. Zhou, *Geometry on probability spaces*, preprint, 2008.

[4] M. Belkin and P. Niyogi, *Laplacian eigenmaps for dimensionality reduction and data representation*, Neural Comput. **15** (2003), 1373–1396.

Online Learning Algorithms as Stochastic Approximations of the Regularization Path

Pierre Tarrès
(joint work with Yuan Yao)

Consider the following classical problem of learning from examples: given a sequence of i.i.d. random samples $(z_t = (x_t, y_t))_{t\in\mathbb{N}}$ drawn from a probability measure ρ on $X \times Y$, one seeks to approximate the *regression function*

$$f_\rho(x) := \int_Y y d\rho_{Y|x},$$

i.e., the conditional expectation of y given x.

We study here *online learning algorithms,* which are recursive, contrary to *batch learning algorithms* which process the data once and for all at some fixed time m. We show [7], using stochastic approximation techniques, how their convergence rates can match the batch learning ones.

The quality of the estimate one can obtain depends on the regularity of f_ρ, measured through a Mercer kernel $K : X \times X \longrightarrow \mathbb{R}$ (continuous, symmetric and positive semidefinite). The Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}_K$ is defined as the closure of the linear span of the set of functions $\{K_x := K(x, .),\ x \in X\}$, with the inner product, denoted as $< .,. >_K$, satisfying $< K_x, K_y >_K = K(x, y)$.

Recall the reproducing property $< K_x, f >= f(x)$, for all $x \in X$, $f \in \mathcal{H}_K$, which implies in particular that $\|f\|_\infty \leq \kappa\|f\|_K$, where $\kappa := \sup_{x\in X} \sqrt{K(x, x)}$.

We analyze *online* algorithms of the type

$$f_t = f_{t-1} - \gamma_t[(f_{t-1}(x_t) - y_t)K_{x_t} + \lambda_t f_{t-1}], \qquad \text{for some } f_0 \in \mathcal{H}_K, \text{ e.g. } f_0 = 0,$$

with gain sequences $(\lambda_t)_{t\in\mathbb{N}}$ and $(\gamma_t)_{t\in\mathbb{N}}$ taking values in $\mathbb{R}_+ \setminus \{0\}$, originally introduced by Smale and Yao in [5], and further studied by Yao in [8]. The recursion can be interpreted as a stochastic gradient descent

$$f_t = f_{t-1} - \mathsf{grad}\ V_{z_t}^{\lambda_t}(f_{t-1}),$$

where

$$V_z^\lambda(f) := \frac{1}{2}[(f(x) - y)^2 + \lambda\|f\|_k^2]$$

for all $f \in \mathcal{H}_K$, $z \in X \times Y$ and $\lambda \in \mathbb{R}_+$. One of the advantages of such algorithms is their computational complexity, which is quadratic in time in the worst case, and can be linear at the cost of a large memory allocation. In comparison, the batch learning Tikhonov regularization scheme typically involves the inverse of a matrix, which is $O(t^3)$ in the worst case.

We optimize the choice of $(\lambda_t)_{t\in\mathbb{N}}$ and $(\gamma_t)_{t\in\mathbb{N}}$, as a function of the regularity of f_ρ. More precisely, let ρ_X be the induced marginal probability measure from ρ on X, and let $L_K : \mathcal{L}^2(\rho_X) \longrightarrow \mathcal{L}^2(\rho_X)$ be the self-adjoint operator defined by

$$L_K(f)(x) = \int_X K(x, y)f(y)d\rho_X(y) =< K_x, f >_{\mathcal{L}^2(\rho_X)}, \quad x \in X,$$

which is positive and compact, so that we can define (through any orthornomal system), the operators $L_K^r : \mathcal{L}^2(\rho_X) \longrightarrow \mathcal{L}^2(\rho_X)$ for all $r \in \mathbb{R}_+$.

Assume that f_ρ lies in the image of L_K^r. We show that, if we choose $f_0 := 0$, and

$$\gamma_t := a(t+t_0)^{-\frac{2r}{2r+1}},\ \lambda_t := b(t+t_0)^{-\frac{1}{2r+1}}, \tag{1}$$

for some $t_0 := \mathsf{Cst}(\kappa)$, a, $b := \mathsf{Cst}(M_\rho, \|L_K^{-r} f_\rho\|_K)$ then, with confidence $1-\delta$,

$$\|f_t - f_\rho\|_K \leq \mathsf{Cst}(\kappa, M_\rho, \|L_K^{-r} f_\rho\|_{\mathcal{L}^2(\rho_X)}) \left(\log \frac{2}{\delta}\right) t^{-\frac{2r-1}{4r+2}},$$

and

$$\|f_t - f_\rho\|_{\mathcal{L}^2(\rho_X)} \leq \mathsf{Cst}(\kappa, M_\rho, \|L_K^{-r} f_\rho\|_{\mathcal{L}^2(\rho_X)}) \left(\log \frac{2}{\delta}\right)^2 t^{-\frac{r}{2r+1}}.$$

The choice $a = b := 1$ yields the same result, at the expense, however, of the constants involved.

The exponent in t in the $\mathcal{H}_K$-norm rate is the same as the best known one in batch learning, obtained by Smale and Zhou [6], and the mean square distance convergence rate is optimal in the sense that it reaches the minimax and individual lower rates (see for instance Caponnetto and de Vito [2]).

The choice of these gain sequences in (1) is derived from the analysis of the algorithm as a stochastic approximation of a Tikhonov regularization path converging to the regression function.

In the talk we explain some previous results on the convergence rates of stochastic algorithms, in particular the "1/2-phase transition", which also plays an important rôle in the Pólya urn model (see for instance Athreya and Karlin [1] or, more recently, Pouyanne [4]). We show how these finite-dimensional techniques can be extended to the infinite-dimensional online algorithm considered here, using on the one hand some martingale and reverse-martingale expansions, and on the other hand probabilistic exponential inequalities on Banach spaces provided by Pinelis [3].

References

[1] K. B. Athreya and S. Karlin, *Embedding of urn schemes into continuous time Markov branching processes and related limit theorems*, Ann. Math. Statist. **39** (1968), 1801–1817.

[2] A. Caponnetto and E. De Vito, *Optimal rates for regularized least squares algorithm*, Found. Comput. Math. **7**, no. 3 (2007), 331–368.

[3] I. Pinelis, *Optimum bounds for the distributions of martingales in Banach spaces*, Ann. Prob. **22**, no. 4 (1994), 1679–1706.

[4] N. Pouyanne, *An algebraic approach to Pólya processes*, Ann. Inst. H. Poincaré, **44**, no. 2 (2008), 293-323.

[5] S. Smale and Y. Yao, *Online learning algorithms*, Found. Comput. Math. **6**, no. 2 (2006), 145–170.

[6] S. Smale D.-X. Zhou, *Learning theory estimates via integral operators and their approximations*, Constr. Approx. **26**, no. 2 (2007), 153–172.

[7] P. Tarrès and Y. Yao, *Online learning algorithms as stochastic approximations of the regularization path*, Preprint, 2006.

[8] Y. Yao, *On complexity issue of online learning algorithms*, accepted for publication in IEEE Transactions on Information Theory.

Primal and Dual Model Representations in Supervised and Unsupervised Kernel-based Learning

Johan A.K. Suykens

Support vector machine classification and regression problems have been characterized as convex optimization problems. One makes use then of a high dimensional feature map in the primal and expresses the solution in the (Lagrange) dual through a positive definite kernel function. In this talk we explain about the general role that primal and dual model representations may play towards constructive approximation and integrative understanding of kernel-based learning methods.

Many basic problems in supervised and unsupervised learning, including regression, classification, principal component analysis, canonical correlation analysis, spectral clustering, data visualization and others can be understood in terms of simple core models involving a least squares objective and equality constraints. Both in supervised and unsupervised learning problems, the formulations have explicit underlying models which enable to make out-of-sample extensions. This is relevant in model tuning and selection for achieving a good generalization of the model. It also enables making predictions, which is illustrated by examples in spectral clustering and independent component analysis.

Starting from core models different types of constraints can be added. We illustrate this for problems of structure detection, semi-supervised learning, system identification and time-series prediction. The optimal model representations and kernel based model solutions both follow from the conditions for optimality.

For data visualization and dimensionality reduction we present kernel maps with a reference point [1]. Unlike methods as locally linear embedding, Laplacian eigenmaps and diffusion maps which are characterized by eigenvalue problems, kernel maps with a reference point lead to solving linear systems. The underlying model allows to make out-of-sample extensions and cross-validation based learning. It contains an additional regularization term which is a modification to locally linear embedding.

Sparse representations are readily obtained through fixed-size methods. One employs then a finite dimensional approximation to the feature map with estimation in the primal based on a subset of the training data. These methods are suitable for handling large data sets. Finally, aspects of robustness are addressed in relation to learning theory.

Acknowledgements. Research supported by Research Council K.U. Leuven: GOA AMBioRICS, CoE EF/05/006, OT/03/12, PhD/postdoc & fellow grants; Flemish Government: FWO PhD/postdoc grants, FWO projects G.0499.04, G.0211.05, G.0226.06, G.0302.07; Research communities (ICCoS, ANMMM, MLDM); AWI: BIL/05/43, IWT: PhD Grants; Belgian Federal Science Policy Office: IUAP DYSCO. Publications available at http://www.esat.kuleuven.be/scd/.

References

[1] J.A.K. Suykens, *Data Visualization and Dimensionality Reduction using Kernel Maps with a Reference Point*, IEEE Transactions on Neural Networks, in press.

Sparsity in Statistical Learning

Alexandre B. Tsybakov

The aim of talk was to give an introduction to statistical estimation in high-dimensional models (where the dimension p of the vector of unknown parameters is larger than the sample size n) under sparsity scenario. The model is called sparse if the number of non-zero coordinates of the vector of unknown parameters is much smaller than p. The quality of sparse estimation is usually assessed in terms of model selection consistency (i.e., recovering of the set of non-zero coordinates) and *sparsity oracle inequalities* (SOI) for the prediction risk. One of the most important issues is to build methods that attain optimal performances with respect to these two criteria under minimal assumptions on the dictionary (for example, in linear regression, this requirement is translated as minimal assumptions on the design matrix X). Sparse statistical estimation is closely related to the problem of compressed sensing in approximation theory, but is more complex because the noise is added. It is also related to the problem of aggregation of estimators since, using sparse estimation methods obeying the SOI, we can construct aggregates that are simultaneously optimal for convex, linear and model selection type aggregation.

Most popular methods of sparse statistical estimation are mainly of the two types. Some of them, like the BIC, enjoy nice theoretical properties without any assumption on the dictionary but are computationally infeasible starting from relatively modest dimensions p. Others, like the Lasso or the Dantzig selector, are easily realizable for very large p but their theoretical performance is conditioned by severe restrictions on the dictionary. In this talk we focus on *Sparse Exponential Weighting* [4-5], a new method of sparse recovery in regression, density and classification models realizing a compromise between theoretical properties and computational efficiency. The theoretical performance of the method is comparable with that of the BIC in terms of SOI for the prediction risk. No assumption on the dictionary is required when the squared loss is considered. At the same time, the method is computationally feasible for relatively large dimensions p. It is constructed using exponential weighting with suitably chosen priors, and its analysis is based on the PAC-Bayesian ideas in statistical learning. We develop a general technique to derive sparsity oracle inequalities from the PAC-Bayesian bounds. The talk is based on the papers [1-6].

References

[1] P.J. Bickel, Y. Ritov, and A.B. Tsybakov, *Simultaneous analysis of Lasso and Dantzig selector*, Ann. Statist., to appear.

[2] F. Bunea, A.B. Tsybakov, and M.H. Wegkamp, *Aggregation for Gaussian regression*, Ann. Statist., **35** (2007), 1674–1697.

[3] F. Bunea, A.B. Tsybakov, and M.H. Wegkamp, *Sparsity oracle inequalities for the Lasso*, Electronic J. Statist., **1** (2007), 169–194.
[4] A. Dalalyan, and A.B. Tsybakov, *Aggregation by exponential weighting and sharp oracle inequalities*, Proc. 20th Annual Conf. Learning Theory (COLT-2007), Lecture Notes in Artificial Intelligence, **4539** (2007), 97–111, Springer, Berlin-Heidelberg.
[5] A. Dalalyan, and A.B. Tsybakov. *Aggregation by exponential weighting, sharp PAC-Bayesian bounds and sparsity*, Machine Learning, **72** (2008), 39–61.
[6] K. Lounici. *Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators*, Electronic J. Statist., **2** (2008), 90–102.

Analysis of Elastic-Net Regularization

Lorenzo Rosasco
(joint work with Christine De Mol, Ernesto De Vito)

In many learning problems, a major goal besides prediction is that of *selecting the variables* that are *relevant to achieve good predictions*. In the problem of variable selection we are given a set $(\varphi_\gamma)_{\gamma\in\Gamma}$ of functions from the input space $\mathcal{X}$ into the output space $\mathcal{Y}$ and we aim at selecting those functions which are needed to find a good representation of the regression function f^* on the basis of n input-output samples. In last decade many different algorithms have been introduced to solve such problem, such as forward stepwise regression, Lasso and greedy algorithms. However these procedures have drawbacks if there are highly correlated features. To overcome this problem, Zou and Hastie suggest a new method, called the elastic-net regularization [3]. In our work we study several properties of this estimation procedure with the setting of statistical learning (see [2] for details). In particular, we prove consistency for prediction and variable selection under some adaptive and non-adaptive choices for the regularization parameter. As an extension of the setting originally proposed in [3], our setting is random-design regression where we allow the response variable to be vector-valued and we consider prediction functions which are linear combination of elements (*features*) in an infinite-dimensional dictionary. The elastic-net scheme is defined by the minimization of the empirical risk penalized with a (weighted) elastic-net penalty, that is, given a sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ of i.i.d random pairs in $(\mathcal{X}, \mathcal{Y})$, the estimator vector β_n^λ is

$$\beta_n^\lambda = \operatorname*{argmin}_{\beta\in\ell_2} \frac{1}{n}\sum_{i=1}^{n} |Y_i - f_\beta(X_i)|^2 + \lambda \sum_{\gamma\in\Gamma} (w_\gamma|\beta_\gamma| + \varepsilon\beta_\gamma^2)$$

$$f_\beta = \sum_{\gamma\in\Gamma} \beta_\gamma \varphi_\gamma,$$

where $(w_\gamma)_{\gamma\in\Gamma}$ is a family of positive weights enforcing more or less sparsity, λ is a regularization parameter controlling the trade-off between the empirical error and the penalty, and ε is a tuning positive parameter that controls the trade-off between the ℓ_1-penalty (pure Lasso) and the ℓ_2-penalty (regularized least-squares regression). The ℓ_1-penalty has selection capabilities since it enforces sparsity of

the solution, whereas the ℓ_2-penalty induces a linear shrinkage on the coefficients leading to stable solutions.

Under the assumption that the features satisfy $\sup_{x\in\mathcal{X}}\sum_{\gamma\in\Gamma}\|\varphi_{\gamma(x)}\|_{\mathcal{Y}}^2<\infty$ and the noise $Y_i-f^*(X_i)$ has exponential tails, that is,

$$\mathbb{E}\left[\exp\left(\frac{\|Y_i-f^*(X_i)\|_{\mathcal{Y}}}{L}\right)-\frac{\|Y_i-f^*(X_i)\|_{\mathcal{Y}}}{L}-1\Big|X_i\right]\qquad\leq\frac{\sigma^2}{2L^2},$$

we prove that, if the regularization parameter $\lambda=\lambda_n$ satisfies $\lim_{n\to\infty}\lambda_n=0$ and $\lim_{n\to\infty}(\lambda_n\sqrt{n}-2\log n)=+\infty$, then

$$\lim_{n\to\infty}\|\beta_n^{\lambda_n}-\beta^{\varepsilon}\|_2=0\qquad\text{with probability one,}$$

where the vector β^{ε}, which we call the *elastic-net representation* of f^*, is the minimizer of

$$\min_{\beta\in\ell_2}\left(\sum_{\gamma\in\Gamma}w_\gamma|\beta_\gamma|+\varepsilon\sum_{\gamma\in\Gamma}|\beta_\gamma|^2\right)\qquad\text{subject to}\qquad\sum_{\gamma\in\Gamma}\beta_\gamma\varphi_\gamma=f^*.$$

The vector β^{ε} exists and is unique provided that the regression function f^* admits a *sparse representation on the dictionary*, i.e. $f^*=\sum_{\gamma\in\Gamma}\beta_\gamma^*\varphi_\gamma$ for at least a vector $\beta^*\in\ell_2$ such that $\sum_{\gamma\in\Gamma}w_\gamma|\beta_\gamma^*|$ is finite. Notice that, when the features are linearly dependent, there is a problem of identifiability since there are many vectors β such that $f^*=\sum_{\gamma\in\Gamma}\beta_\gamma\varphi_\gamma$. The elastic-net regularization scheme forces $\beta_n^{\lambda_n}$ to converge to β^{ε}. As a consequence of the above convergence result, one easily deduces the consistency of the corresponding prediction function $f_n:=\sum_{\gamma\in\Gamma}(\beta_n^{\lambda_n})_\gamma\varphi_\gamma$, that is, $\lim_{n\to\infty}\mathbb{E}[|f_n-f^*|^2]=0$ with probability one. When the regression function does not admit a sparse representation, we can still prove the previous consistency result for f_n provided that the regression function is bounded and the linear span of the features is dense in $L^2(\mathcal{X},Q,\mathcal{Y})$, where Q is the marginal distribution of X. Both the above convergence results are based on the fact that β_n^{λ} is the fixed point of the following contractive map

$$\beta=\frac{1}{\tau+\varepsilon\lambda}\mathbf{S}_\lambda\left(\tau I-\Phi_n^*\Phi_n)\beta+\Phi_n^*Y\right)\tag{1}$$

where τ is a suitable relaxation constant, $\Phi_n^*\Phi_n$ is the matrix with entries $(\Phi_n^*\Phi_n)_{\gamma,\gamma'}=\frac{1}{n}\sum_{i=1}^n<\varphi_\gamma(X_i),\varphi_{\gamma'}(X_i)>_{\mathcal{Y}}$, Φ_n^*Y is the vector $(\Phi_n^*Y)_\gamma=\frac{1}{n}\sum_{i=1}^n<\varphi_\gamma(X_i),Y_i>_{\mathcal{Y}}$. Moreover, $\mathbf{S}_\lambda(\beta)$ is the soft-thresholding operator acting componentwise as follows

$$[\mathbf{S}_\lambda(\beta)]_\gamma=\begin{cases}\beta_\gamma-\frac{\lambda w_\gamma}{2} & \text{if}\quad \beta_\gamma>\frac{\lambda w_\gamma}{2}\\ 0 & \text{if}\quad |\beta_\gamma|\leq\frac{\lambda w_\gamma}{2}\\ \beta_\gamma+\frac{\lambda w_\gamma}{2} & \text{if}\quad \beta_\gamma<-\frac{\lambda w_\gamma}{2}\end{cases}.$$

As a by-product of (1), β_n^{λ} has only a finite number of non-zero components, corresponding to the features whose weight satisfies $w_\gamma<\frac{C_n}{\lambda}$, where C_n is a known constant. Moreover β_n^{λ} can be computed by means of an iterative algorithm. This

procedure is completely different from the modification of the LARS algorithm used in [3] and is akin instead to the algorithm developed in [1].

Finally, we use a data-driven choice for the regularization parameter, based on the so-called balancing principle, to obtain non-asymptotic bounds which are adaptive to the unknown regularity of the regression function. More precisely, letting $\lambda_k = \lambda_0 q^k$ be a geometric sequence with $q > 1$, we define

$$\lambda_n^+ = \max\{\lambda_k | \|\beta_n^{\lambda_j} - \beta_n^{\lambda_{j-1}}\|_2 \leq \frac{4D}{\sqrt{n}\varepsilon\lambda_{j-1}} \text{ for all } j = 0, \dots, k\},$$

where D is a suitable constant. If β^ε is such that for some unknown $a \in (0, 1)$ it satisfies the a-priori bound

$$\|\beta^\lambda - \beta^\varepsilon\|_2 = O(\lambda^a) \qquad \text{where}$$
$$\beta^\lambda = \operatorname*{argmin}_{\beta \in \ell_2} \mathbb{E}[\|Y - f_\beta(X)\|_{\mathcal{Y}}^2] + \lambda \sum_{\gamma \in \Gamma} (w_\gamma |\beta_\gamma| + \varepsilon \beta_\gamma^2),$$

then we prove that $\|\beta^{\lambda_n^+} - \beta^\varepsilon\|_2 = O(n^{-\frac{a}{2(a+1)}})$.

References

[1] I. Daubechies, M. Defrise, and C. De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Comm. Pure Appl. Math.*, 57(11):1413–1457, 2004.

[2] C. De Mol, E. De Vito, and L. Rosasco. Elastic-Net Regularization in Learning Theory. preprint arXiv:0807.3423 (July 2008)

[3] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B*, 67(2):301–320, 2005.

Spectral Regularization for Multi-task Learning

Massimiliano Pontil

(joint work with Andreas Argyriou, Charles Micchelli, Yiming Ying)

We are interested in the problem of learning multiple regression or classification functions (tasks) simultaneously. We present a method for learning a set of features which are shared across the tasks [1]. The method is based on a non-convex regularizer which encourages the number of such features to be small. We highlight the observation that the method is equivalent to solving a convex optimization problem, for which there is an iterative algorithm. The algorithm has a simple interpretation and converges to an optimal solution.

1. **Notation.** We begin by introducing our notation. We let $\mathbb{R}$ be the set of real numbers and $\mathbb{R}_+$ the subset of nonnegative ones. If $w, u \in \mathbb{R}^d$, we define $\langle w, u\rangle := \sum_{i=1}^d w_i u_i$ and $\|w\|_2 = \sqrt{\langle w, w\rangle}$. If A is a $d \times T$ matrix we denote by $a^i \in \mathbb{R}^T$ and $a_t \in \mathbb{R}^d$ the i-th row and the t-th column of A respectively. We denote by $\mathbf{S}_{++}^d$ the set of symmetric and positive definite matrices. If D is a $d \times d$ matrix, we define $\text{trace}(D) := \sum_{i=1}^d D_{ii}$. If $w \in \mathbb{R}^d$, we denote by $\text{Diag}(w)$ or $\text{Diag}\,(w_i)_{i=1}^d$

the diagonal matrix having the components of vector w on the diagonal. We let $\mathbf{O}^d$ be the set of $d \times d$ orthogonal matrices.

2. **Problem formulation.** We are given T supervised learning tasks. For every $t = 1, \ldots, T$, the corresponding task is identified by a function $f_t : \mathbb{R}^d \to \mathbb{R}$. For each task, we are given a set of m input/output examples

$$(x_{t1}, y_{t1}), \ldots, (x_{tm}, y_{tm}) \in \mathbb{R}^d \times \mathbb{R} \, .$$

We wish to use the available examples in order to uncover *particular* relationships across the tasks. Our working assumption is that the tasks *all share a small set of features*, namely the functions f_t can be represented as a linear combination of a few feature functions. For simplicity, we consider linear homogeneous features, each of which is represented by a vector $u_i \in \mathbb{R}^d$ – extensions to non-linear features are dealt with in [1]. Furthermore, we assume that the vectors u_i are orthogonal and, so, we consider only up to d of such vectors.

If we denote by $U \in \mathbf{O}_d$ the matrix whose columns are the vectors u_i, the task functions can be written as $f_t(x) = \langle u_i, x \rangle = \langle a_t, U^\top x \rangle$, $x \in \mathbb{R}^d$, where $a_t = (a_{t1}, \ldots, a_{td})^\top$ is the vector of regression coefficients for the t-th task.

Our assumption that the tasks share a "small" set of features means that the matrix A has "many" rows which are identically equal to zero and, so, the corresponding features (columns of matrix U) will not be used by any task.

The learning method described in [1] is to solve the optimization problem

$$\min \left\{ \mathbb{E}(A, U) : U \in \mathbf{O}^d, A \in \mathbb{R}^{d \times T} \right\}, \tag{1}$$

$$\mathbb{E}(A, U) = \sum_{t=1}^{T} \sum_{i=1}^{m} L(y_{ti}, \langle a_t, U^\top x_{ti} \rangle) + \gamma \|A\|_{2,1}^2, \tag{2}$$

where $\gamma > 0$ is a regularization parameter.

The first term in (2) is the average of the error across the tasks, measured according to a prescribed loss function $L : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ which we assume to be convex in the second argument. The second term is a regularization function which penalizes the $(2,1)$-norm of matrix A. It is obtained by first computing the 2-norms of the (across the tasks) rows a^i (corresponding to feature i) and then the 1-norm of the vector $(\|a^1\|_2, \ldots, \|a^d\|_2)$. The magnitudes of the components of this vector indicate how important each feature is.

We note, that when $T = 1$, function (2) reduces to the well-known 1-norm regularization problem.

The $(2,1)$-norm above favors a small number of nonzero rows in the matrix A, thereby ensuring that few common features will be learned across the tasks. Of course the number of features learned depends on the value of the parameter γ and it will typically be nonincreasing with γ.

We conclude this section by noting that when matrix U is not learned and we set $U = I_{d \times d}$, problem (1) selects a "small" set of variables, common across the tasks.

3. **Equivalent convex problem.** Solving problem (1) is challenging for two main reasons. First, it is a non-convex problem, although it is separately convex in each of the variables A and U. Secondly, the regularizer $\|A\|_{2,1}^2$ is not smooth, which makes the optimization problem more difficult to solve.

Fortunately, problem (1) can be transformed into an equivalent convex problem. To describe this result, for every $W \in \mathbb{R}^{d\times T}$ with columns w_t and $D \in \mathbf{S}_{++}^d$, we define the function

$$\mathbb{R}(W,D) = \sum_{t=1}^{T}\sum_{i=1}^{m} L(y_{ti}, \langle w_t, x_{ti}\rangle) + \gamma \mathrm{trace}(D^{-1}WW^\top). \tag{3}$$

It is then possible to show that Problem (1) is equivalent to the convex optimization problem

$$\inf\left\{\mathbb{R}(W,D) \;:\; W \in \mathbb{R}^{d\times T},\; D \in \mathbf{S}_{++}^d,\; \mathrm{trace}(D) \le 1\right\}. \tag{4}$$

In particular, any minimizing sequence of problem (4) converges to a minimizer of problem (1)-(2). Moreover, the solutions $(\hat{A}, \hat{U})$ and $(\hat{W}, \hat{D})$ of problems (1) and (4) respectively, are related by the formula

$$(\hat{W}, \hat{D}) = \left(\hat{U}\hat{A},\; \hat{U}\,\mathrm{Diag}\left(\frac{\|\hat{a}^i\|_2}{\|\hat{A}\|_{2,1}}\right)_{i=1}^{d}\hat{U}^\top\right).$$

We refer the reader to [1, Sec. 3] for more information on this observation.

Note that, in problem (4) we have bounded the trace of the matrix D from above, because otherwise the optimal solution would be to simply set $D = \infty$ and only minimize the empirical error term in the right hand side of equation (3).

Returning to the discussion of Section 1 on the $(2,1)$-norm, the rank of the optimal matrix D indicates how many common relevant features the tasks share. Indeed, it is clear from the above discussion that the rank of matrix $\hat{D}$ equals the number of nonzero rows of matrix $\hat{A}$.

4. **Learning algorithm.** We now briefly discuss an algorithm for solving problem (4). The algorithm minimizes a perturbation of the objective function (3), in which a perturbation ϵI is added to the matrix $WW^\top$, appearing in the second term in the r.h.s. of (3), where $\epsilon > 0$ and I is the identity matrix. This perturbation keeps D nonsingular and ensures that the infimum over D is always attained.

The algorithm iterates between two steps, until a convergence condition is met. In the first step, we keep D fixed and minimize over W. This step can be carried out independently across the tasks since the regularizer decouples when D is fixed. More specifically, introducing new variables for $D^{-\frac{1}{2}}w_t$ yields a standard 2-norm regularization problem for each task with the same kernel $K(x,x') = x^\top Dx'$. In the second step, we keep the matrix W fixed, and minimize with respect to D. One can show that partial minimization with respect to D has a closed-form solution given by

$$D_\varepsilon(W) = \frac{(WW^\top + \varepsilon I_d)^{\frac{1}{2}}}{\mathrm{trace}(WW^\top + \varepsilon I_d)^{\frac{1}{2}}}. \tag{5}$$

The above algorithm can be interpreted as alternately performing a supervised and an unsupervised step. In the former step we learn task-specific functions (namely the vectors w_t) using a common representation across the tasks. This is because D encapsulates the features u_i and thus the feature representation is kept fixed. In the unsupervised step, the regression functions are fixed and we learn the common representation.

In [1] an analysis of the above algorithm is provided. In particular, it is shown that, for every $\epsilon > 0$, the algorithm converges to a solution of the corresponding perturbed problem. Moreover, as $\epsilon \to 0$, any limiting points of the sequence of such solutions solves problem (4)).

At last, we note that an extension of the ideas discussed here to the case of Shatten norms and other spectral regularizers is presented in [2].

References

[1] A. Argyriou, T. Evgeniou, and M. Pontil. *Convex multi-task feature learning*, Machine Learning, to appear.

[2] A. Argyriou, C.A. Micchelli, M. Pontil, Y. Ying. *A spectral regularization framework for multi-task structure learning*, Proceedings of NIPS 2007.

Learning Gradients: Precitive Models that Infer Geometry and Statistical Dependence

Sayan Mukherjee

Simultaneous dimension reduction and regression considers the problem of finding directions that are informative with respect to predicting the response variable. These methods can be summarized by three categories:

(1) methods based on inverse regression,
(2) methods based on gradients of the regression function,
(3) methods based on combining local classifiers.

Our focus is on the supervised problem however we will use the idea of local estimates that is central to manifold learning.

The first main results in this paper are precise statistical relations between the three approaches. We will show that the gradient estimate is central to this analysis. Our second main result is the inference of graphical models based on gradient estimates. We provide rates of convergence of the estimated graphical model to its population counterpart. These rates and the underlying modeling depend not on the sparsity of the graph but on the rank of the conditional independence matrix or the intrinsic dimension of the gradient on the manifold supporting the data.

The problem of regression can be summarized as estimating the regression function

$$f_r(x) = \mathbb{E}(Y|X = x)$$

from data $D = \{L_i = (Y_i, X_i)\}_{i=1}^n$ where X_i is a vector in a p-dimensional compact metric space $X \in \mathcal{X} \subset \mathbb{R}^p$ and $Y_i \in \mathbb{R}$ is a real valued output. Typically the data

are drawn i.i.d. from a joint distribution, $L_i \overset{i.i.d.}{\sim} \rho(X,Y)$. When p is large the response variable Y may depend on a few directions in $\mathbb{R}^p$,

$$Y = f_r(X) + \varepsilon = g(b_1^T X, \ldots, b_d^T X) + \varepsilon, \tag{1}$$

where ε is noise and $B = (b_1^T, ..., b_d^T)$ is the effective dimension reduction (EDR) space. In this case dimension reduction becomes the central problem in finding an accurate regression model. In the following we develop a theory relating the gradient of the regression function to the above model of dimension reduction.

The central concept of this paper is the gradient outer product matrix. Assume the regression function $f_r(x)$ is smooth, the gradient is given by

$$\nabla f_r = \left(\frac{\partial f_r}{\partial x^1}, ..., \frac{\partial f_r}{\partial x^p} \right)^T$$

and the the gradient outer product matrix Γ is a $p \times p$ matrix with elements

$$\Gamma_{ij} = \left\langle \frac{\partial f_r}{\partial x^i}, \frac{\partial f_r}{\partial x^j} \right\rangle_{L^2_{\rho_X}}, \tag{2}$$

where ρ_X is the marginal distribution of the explanatory variables. Using the notation $a \otimes b = ab^T$ for $a, b \in \mathbb{R}^p$, we can write

$$\Gamma = \mathbb{E}(\nabla f_r \otimes \nabla f_r).$$

The relation between the gradient outer product matrix and dimension reduction is illustrated by the following observation.

Lemma 1. *Under the assumptions of the semi-parametric model* (1)*, the gradient outer product matrix* Γ *is of rank at most* d*. Denote by* $\{v_1, \ldots, v_d\}$ *the eigenvectors associated to the nonzero eigenvalues of* Γ *the following holds:*

$$\operatorname{span}(B) = \operatorname{span}(v_1, \ldots, v_d) \ .$$

The linear regression problem is often stated as

$$y = \beta^T x + \varepsilon, \qquad \mathbb{E}(\varepsilon) = 0. \tag{3}$$

For this model the following relation between gradient estimates and the inverse regression holds.

Proposition 1. *Suppose* (3) *holds. Given the covariance of the inverse regression,* $\Omega_{X|Y} = \operatorname{cov}_Y(\mathbb{E}_X(X|Y))$*, the variance of the output variable,* $\sigma_Y^2 = \operatorname{var}(Y)$*, and the covariance of the input variables,* $\Sigma_X = \operatorname{cov}(X)$*, the gradient outer product matrix is*

$$\Gamma = \sigma_Y^2 \left(1 - \tfrac{\sigma_\varepsilon^2}{\sigma_Y^2}\right)^2 \Sigma_X^{-1} \Omega_{X|Y} \Sigma_X^{-1}, \tag{4}$$

assuming that Σ_X *is full rank.*

In order to generalize Proposition 1 to the nonlinear regression setting we first consider piecewise linear functions. Suppose there exists a non-overlapping partition of the input space

$$\mathcal{X} = \bigcup_{i=1}^{I} R_i$$

such that in each region R_i the regression function f_r is linear

$$f_r(x) = \beta_i^T x + \varepsilon_i,\ \mathbb{E}(\varepsilon_i) = 0 \quad \text{for } x \in R_i. \tag{5}$$

The following corollary is true.

Corollary 1. *Given partitions R_i of the input space for which (5) holds with $\mathbb{E}(\varepsilon_i) = 0$, define in each partition R_i the following local quantities: the covariance of the input variables $\Sigma_i = \mathrm{cov}(X \in R_i)$, the covariance of the inverse regression $\Omega_i = \mathrm{cov}(\mathbb{E}(X \in R_i|Y))$, the variance of the output variable $\sigma_i^2 = \mathrm{var}(Y|X \in R_i)$. Assuming that the matrices Σ_i are full rank, the gradient outer product matrix can be computed in terms of these local quantities*

$$\Gamma = \sum_{i=1}^{I} \rho_X(R_i)\sigma_i^2 \left(1 - \tfrac{\sigma_{\varepsilon_i}^2}{\sigma_i^2}\right)^2 \Sigma_i^{-1}\,\Omega_i\,\Sigma_i^{-1}, \tag{6}$$

where $\rho_X(R_i)$ is the measure of partition R_i with respect to the marginal distribution ρ_X.

A natural idea in multivariate analysis is to model the conditional independence of a multivariate distribution using a graphical model over undirected graphs. The theory of Gauss-Markov graphs was developed for multivariate Gaussian densities

$$p(x) \propto \exp\left(-\frac{1}{2}x^T JX + h^T x\right),$$

where the covariance is J^{-1} and the mean is $\mu = J^{-1}h$. The result of the theory is that the precision matrix J, given by $J = \Sigma_X^{-1}$, provides a measurement of conditional independence. The meaning of this dependence is highlighted by the partial correlation matrix R_X where each element R_{ij} is a measure of dependence between variables i and j conditioned on all other variables $S^{/ij}$ and $i \neq j$

$$R_{ij} = \frac{\mathrm{cov}(x_i, x_j|S^{/ij})}{\sqrt{\mathrm{var}(x_i|S^{/ij})}\sqrt{\mathrm{var}(x_j|S^{/ij})}}.$$

The partial correlation matrix is typically computed from the precision matrix J

$$R_{ij} = -J_{ij}/\sqrt{J_{ii}J_{jj}}.$$

In the regression and classification framework inference of the conditional dependence between explanatory variables has limited information. Much more useful would be the conditional dependence of the explanatory variables conditioned on variation in the response variable. We have shown that both the covariance of

the inverse regression as well as the gradient outer product matrix provide estimates of the covariance of the explanatory variables conditioned on variation in the response variable. Given this observation the inverses of these matrices

$$J_{X|Y} = \Omega_{X|Y}^{-1} \text{ and } J_\Gamma = \Gamma^{-1},$$

provide evidence for the conditional dependence between explanatory variables conditioned on the response. We focus on the inverse of the gradient outer product matrix in this paper since it is of use for both linear and nonlinear functions.

Our proof of the convergence of the estimated conditional dependence matrix $(\hat{\Gamma})^{-1}$ to the population conditional dependence matrix Γ^{-1} relies on the assumption that the gradient outer product matrix being low rank. This again highlights the difference between our modeling assumption of low rank versus sparsity of the conditional dependence matrix. Since we assume that both Γ and $\hat{\Gamma}$ are singular and low rank we use pseudo-inverses in order to construct the dependence graph.

Proposition 2. *Let Γ^{-1} be the pseudo-inverse of Γ. Let the eigenvalues and eigenvectors of $\hat{\Gamma}$ be $\hat{\lambda}_i$ and $\hat{v}_i$ respectively. If $\varepsilon > 0$ is chosen so that $\varepsilon = \varepsilon_n = o(1)$ and $\varepsilon_n^{-1}\|\hat{\Gamma} - \Gamma\| = o(1)$, then the convergence*

$$\sum_{\hat{\lambda}_i > \varepsilon} \hat{v}_i \hat{\lambda}_i^{-1} \hat{v}_i \to \Gamma^{-1}$$

holds in probability.

High Dimensional Learning via Sparse Occupancy Trees

Peter Binev

(joint work with Wolfgang Dahmen and Ronald DeVore)

Let X be a set of points in $\mathcal{X} \subset \mathbb{R}^d$ and let for each $x_j \in X$ we are given the computed value $y_j \in [-M, M]$ of a function at x_j. Our goal is to find an approximation to this function at any query point from $\mathcal{X}$. We consider the points (x, y) as random drawings from an unknown probability measure ρ on $\mathcal{X} \times [-M, M]$. Then the function of interest is the regression function $f_\rho(x)$ which is defined as the expected value of y given x. We focus here on problems in very high dimension d, in which typically the number of data points $m := \#X$ is significantly less than 2^d. In particular, this means that the function is severely undersampled, if ρ has full dimensionality. Standard methods suitable for low dimensions usually do not apply well in these settings, since often some parameters exhibit exponential dependence on the dimension. This effect is called sometimes 'curse of dimensionality'. In certain situations it is possible to avoid it and one of them is the case in which the measure ρ is concentrated around a set of low dimensionality. The question is how to design a method which takes advantage of it.

We need three basic ingredients to develop adaptive methods for solving such problems:

(i) a data structure that allows an effective analysis and fast calculation of the approximation at any query point;
(ii) a framework allowing development of fast adaptive algorithms with guarantied near-best performance;
(iii) theoretical analysis and estimates with high probability of the approximation error.

We consider a fixed procedure for receiving adaptive partitions of the domain $\mathcal{X}$ and build the corresponding master tree $\mathcal{T}$ describing it. Based on the finest resolution $\epsilon > 0$ of the data, we prescribe the maximal depth of $\mathcal{T}$ to be in the range of $|\log \epsilon|$. In order to develop an effective data structure for our algorithms we propose a special organization of the data which we call *sparse occupancy trees*. For each point $x \in X$ we calculate a bitstream showing the sequence of splits up to the leaf node Δ in $\mathcal{T}$ to which x belongs. The lexicographical order of the points in X with respect to these bitstreams allows a fast access to any (occupied) node of the tree $\mathcal{T}$ which is represented by an interval in this list. This structure allows the development of fast algorithms for analysis and fast evaluation. We create sparse occupancy trees $T = T(X)$ based on subtrees of $\mathcal{T}$ consisting of occupied nodes. In addition, the edges from a parent node to a *single* child node are collapsed and these nodes are merged. We can define sequences of sparse occupancy trees T_j starting with T_0 which contains only the root node of $\mathcal{T}$ and then subdividing consecutively some leaf node of T_j to receive the next tree T_{j+1}. The complexity of a tree T will be measured by the number of subdivisions $\mathcal{N}(T)$ needed to receive T from T_0, or equivalently the number of its internal vertices. The corresponding partition of $\mathcal{X}$ generated by a tree T will be denoted by $\Lambda = \Lambda(T)$.

Next we develop an approximation of the empirical data $\mathcal{Z} = \{(x, y)\}$ on partitions $\Lambda(T)$. First, we set $e_{\mathcal{Z}}(\Delta)$ to be the empirical error of the approximation at the points from X attached to the node $\Delta \in \mathcal{T}$. In case of piecewise constant approximation this error is

$$e_{\mathcal{Z}}(\Delta) := \inf_{y \in \mathbb{R}} \left(\frac{1}{\#\{x_i \in \Delta \cap X\}} \sum_{x_i \in \Delta \cap X} |y_i - y|^2 \right).$$

For a given tree T and its corresponding partition $\Lambda(T)$ we set the global error to be

$$\mathcal{E}_{\mathcal{Z}}(T) = \mathcal{E}_{\mathcal{Z}}(\Lambda) := \sum_{\Delta \in \Lambda} e_{\mathcal{Z}}(\Delta).$$

Then the best approximation σ_n on partitions of complexity n is defined as

$$\sigma_n(\mathcal{T}) = \sigma_n(f_\rho)_{\mathcal{Z}} := \inf_{T : \mathcal{N}(T) \le n} \mathcal{E}\mathcal{Z}(T) \ .$$

In [2] we have considered near-best algorithms for general type of the errors. To establish similar results for sparse occupancy trees we require that the following subadditivity condition holds

$$e_{\mathcal{Z}}(\Delta) \ge \sum_{\Delta' \in \mathcal{C}(\Delta)} e_{\mathcal{Z}}(\Delta') \tag{1}$$

for any node $\Delta \in \mathcal{T}$ and the set of its children $\mathcal{C}(\Delta)$. Using a modification of the methods from [2] and the particular scheme from [1] we develop an adaptive tree algorithm for building a near-best sparse occupancy tree and prove the following result.

Theorem 1. *Let the errors $e(\Delta)$ in $\mathcal{T}$ satisfy the subadditivity condition (1). Then at each step of the* adaptive tree algorithm *the output tree T satisfies*

$$\mathcal{E}_{\mathcal{Z}}(T) \leq \left(\frac{\mathcal{N}(T)+1}{\mathcal{N}(T)-n+1}\right)\ \sigma_n(\mathcal{T})$$

whenever $n \leq \mathcal{N}(T)$.

Note that the above estimate does not depend directly on the number of occupied cells $\#\Lambda$. To connect with the next result we have to require that $\#\Lambda(T) \sim \mathcal{N}(T)$. This is true in particular in case $\mathcal{T}$ is a binary tree, or more generally, if the number of children of the nodes in $\mathcal{T}$ is bounded by a fixed (small) constant.

To prove that our empirical solution gives good approximation to the regression problem, we have to establish a relationship between the empirical error and the actual error. We consider piecewise constant approximation on *sparse* partitions Λ received via the adaptive procedure established by $\mathcal{T}$. 'Sparse' means that it is not necessary for the cells $\Delta \in \Lambda$ to cover the entire domain $\mathcal{X}$. The error of approximation on Λ is defined as

$$\mathcal{E}(\Lambda) := \int_{\mathcal{X}} \left| f_\rho(x) - \sum_{\Delta\in\Lambda} \mathbf{1}_\Delta(x)\mathbb{E}(y|x\in\Delta)\right|^2 d\rho_{\mathcal{X}}\ ,$$

where $\mathbf{1}_\Delta$ is the indicator function for the set Δ, $\mathbb{E}(y|x\in\Delta)$ is the conditional expectation for y given that $x \in \Delta$, and $\rho_{\mathcal{X}}$ denotes the marginal probability measure on $\mathcal{X}$. The best approximation is defined in terms of the complexity of the sparse partition Λ

$$\sigma_n(f_\rho) := \inf_{\Lambda:\#\Lambda\leq n}\ \mathcal{E}(\Lambda)\ .$$

Using the standard techniques from Learning Theory (see e.g. [3]) we can establish the following relationship between $\mathcal{E}(\Lambda)$ and its empirical counterpart $\mathcal{E}_{\mathcal{Z}}(\Lambda)$.

Theorem 2. *Given $q \in (0,1)$ and an upper bound $N \geq \#\Lambda$, for every $\eta > 0$ we have*

$$|\mathcal{E}(\Lambda) - \mathcal{E}_{\mathcal{Z}}(\Lambda)| \leq \eta^2 + q\mathcal{E}(\Lambda)$$

with probability at least $1 - 6e^{-\frac{m\eta^2}{(48+24/q)NM^2}}$.

Choosing $q = \frac{1}{3}$ and requiring that $\mathcal{E}(\mathbf{z}) \geq 3\eta^2$, we receive from the above theorem that with high probability (at least $1 - 6e^{-\frac{m\eta^2}{120NM^2}}$) the empirical error and the actual error differ at most by a factor of two

$$\frac{1}{2} \leq \frac{\mathcal{E}(\Lambda)}{\mathcal{E}_{\mathcal{Z}}(\Lambda)} \leq 2\ .$$

It is customary to set the probability for such estimates in the form $1 - m^{-\beta}$ for some (large) positive β. Then we can determine the value of η setting

$$\eta^2 = \frac{NM^2}{m} \left(\beta \ln m + \ln 6\right) \left(48 + \frac{24}{q}\right) .$$

This also gives a natural stopping criterion for the tree algorithm which written in a compact form is as follows:

Grow the tree T until for the corresponding partition $\Lambda = \Lambda(T)$ the inequality

$$\frac{\mathcal{E}_{\mathcal{Z}}(\Lambda)}{\#\Lambda} \geq C_0 M^2 \frac{\ln m}{m}$$

holds with a given absolute constant C_0.

Combining the above results, we can prove the near-best approximation property with high probability showing the instance optimality of our procedure.

Theorem 3. *Given a fixed $\beta > 0$ there exist absolute constants $C_0 = C_0(\beta)$, C_1, and C_2 such that*

$$|f_\rho - f_{\Lambda,\mathbf{z}}| \leq C_1 \sigma_{C_2 \#\Lambda}(f_\rho)$$

with probability at least $1 - m^{-\beta}$.

References

[1] Peter Binev, *Adaptive Methods and Near-Best Tree Approximation*, Mathematisches Forschungsinstitut Oberwolfach, Report No. **29** (2007), 1669–1673.

[2] P. Binev and R. DeVore, *Fast Computation in Adaptive Tree Approximation*, Numer. Math. **97** (2004), 193–217.

[3] P. Binev, A. Cohen, W. Dahmen, R. DeVore, and V. Temlyakov, *Universal algorithms for learning theory - Part I: piecewise constant functions*, Journal of Machine Learning Research **6** (2005), 1297–1321.

Approximate Varieties and Dimension Reduction

Tomas Sauer

Though residing in a high dimensional space, a finite set $\Xi \subset \mathbb{R}^N$ of measured data can nevertheless be generated according to a "simple rule", which could be expressed by the data lying on a lower dimensional manifold. The simplest example would be that some of the variables are irrelevant so that the data is contained in the hyperplane $x_j = c$ for some $j \in \{1, \ldots, N\}$.

Thus, the task of performing *dimension reduction* can be reduced to detecting such a manifold. One particular class of manifolds to be considered are *algebraic varieties*, that is, the common zeros of a finite set F of polynomials in $\Pi = \mathbb{R}[x_1, \ldots, x_N]$. Besides offering a well–defined and well–studied class of manifolds that contains planes and coordinate spaces as above, algebraic surfaces have the advantage of being computationally accessible. Indeed, a variety V can be efficiently described by giving a basis for the associated radical ideal I such that

$$V = \{x \,:\, f(x) = 0,\, f \in I\} ,$$

and polynomials as well as rational functions on an algebraic variety can be handled by means of the isomorphism $\mathbb{R}[V] = \Pi/I$. This allows, for example, for the definition and computational handling of polynomial kernels on the variety.

Naively, the problem could be described as

Find an ideal I such that the associated variety is Ξ.

Polynomial ideals are usually represented by a *basis*, i.e., a finite set F of polynomials such that

$$I = \langle F \rangle = \left\{ \sum_{f \in F} g_f \, f \ : \ g_f \in \Pi \right\},$$

and so the task consists of finding a basis F such that Ξ is the set of solutions of the system $F(x) = 0$. Indeed, the solution to this problem is well–known, namely the Buchberger–Möller–algorithm [1] which even computes a *Gröbner basis* for the ideal I whose associated variety is Ξ. In other words, Ξ itself is already an algebraic variety, but a zero dimensional one that consists of isolated points – which is definitely not what helps in terms of dimension reduction.

In fact, the "right" approach is to look for a "simple" variety V such that $\Xi \subset V$. The intuition behind this concept is clear: if, for example, all points of Ξ lie on an straight line, then the associated variety should be this straight line. Simple varieties are varieties whose ideal is generated by polynomials of a low degree, so the task is finally as follows:

Find, if possible, a set F of low degree polynomials such that $F(\Xi) = 0$.

The key, of course, is "low degree". Since $F(\Xi) = 0$, it is clear that F has to be included in the radical $I(\Xi)$ of all polynomials vanishing at Ξ, and so the "candidates" for F are the low degree polynomials in $I(\Xi)$. These, however, can be conveniently found by considering an H–basis of $I(\Xi)$ and simply picking the low degree elements from this finite set. Recall that an H–basis H is a basis of an ideal I such that any ideal element f can be written as

$$f = \sum_{h \in H} g_h \, h, \qquad \deg f \geq \deg g_h + \deg h,$$

where "deg" denotes the usual total degree. H–bases were introduced by Macaulay in 1914, long before the appearance of Gröbner bases, and they have the advantage that they can be defined and computed in a completely homogeneous way without the need for any artificial term order, cf. [2]. This type of H–bases is less sensitive to perturbations and more useful for numerical ideal computations.

And "numerical" is finally the keyword as we cannot expect Ξ to be *exact* data that would be suitable for a symbolic treatment as in usual computer algebra systems: an arbitrary small perturbation of points on a line would result in point set that is no more on the line and, even worse, that is not located on any simple variety any more. Hence, we have to look for an ideal basis such that $F(\Xi)$ is not necessarily equal to zero but "only" very small, of course after a proper normalization of the coefficients of the polynomials in F. This is the concept of

an *approximate ideal* for a given variety and, conversely, the idea of *approximate variety* associated to a certain ideal. The numerical task of finding an approximate low degree ideal for Ξ again benefits strongly from the concept of H–bases as it permits the use of orthogonality and thus of numerically stable methods to compute polynomials such that $\|F(\Xi)\|_\infty \leq \varepsilon$ for a given $\varepsilon > 0$.

The details and algorithms for such a computation, together with some error estimates can be found in [3].

References

[1] H. M. Möller and B. Buchberger, *The construction of multivariate polynomials with preassigned zeros*, Lecture Notes in Computer Science **144** (1982), 24–31.

[2] H. M. Möller and T. Sauer, *H–bases for polynomial interpolation and system solving*, Advances Comput. Math. **12** (2000), 335–362.

[3] T. Sauer, *Approximate varieties, approximate ideals and dimension reduction*, Numer. Algo. **45** (2007), 295–313.

The LASSO-Patternsearch Algorithm: Finding "Patterns in a Haystack"

Grace Wahba

(joint work with with Weiliang Shi, Steve Wright, Kristine Lee, Ronald Klein and Barbara Klein)

The LASSO-Patternsearch algorithm is proposed to efficiently identify patterns of multiple dichotomous risk factors for dichotomous outcomes of interest in demographic and genomic studies. Briefly, a data set $\{y_i, x(i)\}, i = 1, \cdots, n$ is observed, where i indexes subjects or objects, and $x(i)$ is a long bit sequence associated with subjects or objects, the elements of which are indicator variables for the state of some attribute. y_i is also an indicator variable of some response, or condition. Given this training set, it is desired to build a model which will predict y for new observations on x, and/or to understand the relationships between y and x. The patterns are functions of subsets of bit sequences, and are those that arise naturally from the log linear expansion of the multivariate Bernoulli density. The method is designed for the case where there is a possibly very large number of candidate patterns but it is believed that only a relatively small number are important. Altogether if x is of dimension p there are 2^p potential patterns, although for large p only patterns involving a small number of bits are considered. A LASSO (ℓ_1 penalized log likelihood) is used to greatly reduce the number of candidate patterns, using a novel computational algorithm that can find the global optimum of the associated mathematical programming problem with an extremely large number of unknowns. The patterns surviving the LASSO are further pruned in the framework of (parametric) generalized linear models. A novel tuning procedure based on the GACV for Bernoulli outcomes, modified to act as a model selector, is used at both steps. We applied the method to myopia data from the population-based

Beaver Dam Eye Study, exposing physiologically interesting interacting risk factors. We then applied the method to data from a generative model of Rheumatoid Arthritis based on Problem 3 from the Genetic Analysis Workshop 15, successfully demonstrating its potential to efficiently recover higher order patterns from attribute vectors of length typical of genomic studies. The paper [1] has recently been published and is available open source on the web at the journal website. Other work can be found on the authors' website:
`http://www.stat.wisc.edu/~wahba`, including the very recent report [2].

References

[1] W. Shi, S. Wright, G. Wahba, K. Lee, B. Klein and R. Klein *LASSO-Patternsearch algorithm with application to ophthalmology and genomic data* , Statistics and Its Interface **1** (2008), 137-153.

[2] H. Corrada Bravo, K. Eng, S. Keles, G. Wahba and S. Wright, *Estimating Tree-Structured Covariance Matrices via Mixed Integer Programming with an Application to Phylogenetic Analysis of Gene Expression* , University of Wisconsin-Madison Statistics Dept. TR 1142, July 2008.

On Universal Estimators in Learning Theory

Vladimir Temlyakov

This talk addresses a problem of constructing and analysing estimators for the regression problem in supervised learning. Recently, there has been a big interest in studying universal estimators. Universal means that the estimator does not depend on an a priori assumption on the regression function f_ρ belonging to some class F from a collection of classes $\mathcal{F}$, and the method provides the estimation error for the f_ρ close to the optimal error from the class F. The talk is an illustration of how the general technique of construction of universal estimators, developed in the previous author's paper, can be applied in concrete situations.

A setting of the problem discussed here has been motivated by a recent paper by Smale and Zhou. The starting point for us is a given kernel $K(x,u)$ defined on $X \times \Omega$. On the base of this kernel we build an estimator that is universal for classes defined in terms of nonlinear approximations with regard to the system $\{K(\cdot,u)\}_{u\in\Omega}$. We apply the Relaxed Greedy Algorithm for construction of an estimator that is easily implemented.

A Poor Man's Wilks Phenomenon

Stéphane Boucheron
(joint work with Pascal Massart)

The Wilks phenomenon appears in parametric statistics when analyzing maximum likelihood density estimation in a regular d-dimensional model. A theorem due to Wilks asserts that both $2(\ell_n(\hat{\theta}) - \ell_n(\theta))$ and $2nD(P_\theta, P_{\hat{\theta}})$ converge in distribution

toward χ^2_d where $\ell_n(\theta)$ denotes the log-likelihood of the sample under the probability indexed by θ, $\hat{\theta}$ the maximizer of the likelihood, and D the relative entropy between two probability laws. This observation is at the root of several developments in model selection (including the development of Akaike's AIC criterion). Here, we consider it as a statement on empirical processes. The sum of the excess empirical risk $2(\ell_n(\hat{\theta}) - \ell_n(\theta))$ and of the excess risk $2nD(P_\theta, P_{\hat{\theta}})$ is equal to the increment of the centered likelihood process between θ (the sampling probability) and $\hat{\theta}$. We wonder whether this increment of a centered empirical process between a fixed and a random point exhibits a non-trivial yet understandable behavior in other settings. The classical proof of the Wilks theorem is based on the asymptotic normality of the MLE estimator and on the Delta-method. Such tools are pointless in statistical learning theory.

If we consider model selection methods based on penalized contrast optimization, in a variety of situations, the expected values of the empirical excess risk $2(\ell_n(\hat{\theta}) - \ell_n(\theta))$ can be considered as a lower bound on the minimum expected values of penalties [1, 2]. This is an incentive to investigate the concentration properties of the excess empirical risk for contrast optimization problems where the ingredients of the proof of the Wilks theorem do not hold.

We will not attempt to establish asymptotic distributional results, but rather focus on possible non-asymptotic statements. It is useful to keep in mind that we intend to investigate the behavior of models of large dimensions. In that case, if some kind of Wilks phenomenon shows up, it is likely that the excess empirical risk will behave according to a Gamma distribution with a large shape parameter and a small rate parameter. This would then imply that the q-norms of the centered empirical excess risk grew like $c\sqrt{Vq} + c'q$ where c and c' are universal constants and V is an upper-estimate of the variance of the empirical excess risk.

We investigate this question for bounded contrast minimization problems. In Statistical Learning, $\mathcal{X} \times \mathcal{Y}$ is endowed with a probability distribution P, the coordinate projections are denoted by X and Y (in binary classification, $\mathcal{Y} = \{-1, 1\}$). A loss function ℓ maps $\mathcal{Y} \times \mathcal{Y}$ on $\mathbb{R}_+$ or rather a bounded interval.

The learning problem consists in finding a function f on $\mathcal{X}$ such that the risk $R(f) = P\ell(f(X), Y)$ is as small as possible. A learning algorithm starts from a sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ collected from independent drawings according to P. A function f^* is assumed to minimize $R(f)$ among all functions such that $(x, y) \mapsto \ell(f(x), y)$ is measurable. A learning algorithm typically looks for a good approximation $\hat{f}$ of f^* in some class $\mathcal{F}$ of classifiers by minimizing an empirical contrast like $R_n(f) = P_n\ell(f(X), Y)$. Thus, a learning task is defined by a sampling probability P over $\mathcal{X} \times \mathcal{Y}$, a loss function ℓ, and a collection of functions $\mathcal{F}$. $\bar{f}$ (respectively $\hat{f}$) denotes a minimizer of R (respectively $R_n(f)$) in $\mathcal{F}$. For a long time, learning theory has been concerned with $R(\hat{f}) - R_n(\hat{f})$ but more recently a great deal of efforts have been dedicated to the analysis of the excess risk $R(\hat{f}) - R(\bar{f})$ (see [6, 8, 9, 3] and references therein). Here, we attempt to build on the recent progress concerning excess risk and recent concentration inequalities due to

the authors [4] and focus on another rather unusual part of the risk functionals $R_n(\bar{f}) - R_n(\hat{f})$.

The complexity assumption aims at describing the richness of the L_2 neighborhood of $\ell(\bar{f}(\cdot), \cdot)$ in the loss class $\mathcal{H}$. There exists some positive sub-linear function ψ such that for all $r \geq r_{\mathrm{cr}}$:

$$\sqrt{n}\mathbb{E}\left[\sup_{h \in \mathcal{H}: P(h-\bar{h})^2 \leq r^2} \left|(P_n - P)(h - \bar{h})\right|\right] \leq \psi(r)\,.$$

The possibility to derive interesting bounds (fast rates) for the excess risk critically relies on the following assumption: There exists some positive sub-linear function ω such that for all $f \in \mathcal{F}$:

$$P\left(\ell(f(X), Y) - \ell(f^*(X), Y)\right)^2 \leq \omega^2\left(\sqrt{R(f) - R(f^*)}\right)\,.$$

The positive root r_* of the equation

$$\sqrt{n}r^2 = \psi(\omega(r))$$

is the key quantity that shows up in the analysis of empirical risk minimization

The positive root r_* of equation $\sqrt{n}r^2 = \psi(\omega(r))$ also shows up in the detailed analysis of the unnormalized empirical excess risk

$$Z = n(R_n(\bar{f}) - R_n(\hat{f}))$$

In the Statistical Learning Theory framework, the latter quantity is the analogue of the likelihood ratio statistics encountered in the density estimation framework. As a function of many independent random variables that do not (should not) depend too much on any of them, Z is likely to be concentrated around its expected value. But the complicated definition of the empirical excess risk makes it non-trivial to identify the scale of this concentration. On the other hand, it is easily recognized that Z is the supremum of an empirical process and deserves to be processed using the tools from empirical process theory.

Upper-bounds on the variance of the empirical excess risk $n(R_n(\bar{f}) - R_n(\hat{f}))$ are first derived by building on the Efron-Stein inequalities and the tail inequalities on the excess risk and the excess empirical risk.

In order to state the results in a concise manner and to emphasize that those results fit in the theory of empirical processes, it is convenient to introduce the loss class $\mathcal{H}$ associated with a model $\mathcal{F}$. To each $f \in \mathcal{F}$ corresponds a loss function h from $\mathcal{U} = \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}_+$, $h((x, y) = \ell(f(x), y)$. Obviously, $Ph = R(f)$ and $P_n h = R_n(f)$. The loss function associated with f^* (respectively $\bar{f}$) is denoted by h^* (respectively $\bar{h}$.) Henceforth, $\hat{h}_n$ denotes the loss function associated with the minimizer of the empirical risk over a sample of size n.

Taking advantage of the fact that the excess empirical risk $nP_n(\bar{h} - \hat{h}_n)$ is the supremum of a bounded empirical process, invoking the Efron-Stein inequalities, it is possible to relate the variance of the empirical excess risk with the expected value of some interesting quantities like the L_2 distance between $\bar{h}$ and $\hat{h}_n$ or the

value of increment of the centered empirical risk process between $\bar{h}$ and $\bar{h}_n$:

$$\begin{aligned}
&\mathrm{Var}\left[nP_n(\bar{h}-\widehat{h}_n)\right] \\
&\quad\leq\quad 2n\left(\mathbb{E}\left[P_n(\bar{h}-\widehat{h}_n)^2\right]+\mathbb{E}\left[P(\bar{h}-\widehat{h}_n)^2\right]\right)
\end{aligned}$$

and

$$\begin{aligned}
&\mathrm{Var}\left[nP_n(\bar{h}-\widehat{h}_n)\right] \\
&\quad\leq\quad 2n\mathbb{E}\left[\left((P_{n-1}-P)(\bar{h}-\widehat{h}_{n-1})\right)\right]+n2\mathbb{E}\left[P\left(\bar{h}-\widehat{h}_n\right)^2\right]
\end{aligned}$$

The second inequality can be immediately exploited using known tail bounds for the excess risk and the empirical excess risk. Indeed, it is now well-known [11, 6, 8, 9, 3] that there exists some universal constants κ_1, κ_2 such that with probability larger than $1-2\delta$:

$$\begin{aligned}
&\max\left(R(\widehat{f})-R(\bar{f}), R_n(\bar{f})-R_n(\widehat{f})\right) \\
&\quad\leq\quad \kappa_1 L(\bar{f})+\kappa_2 r_*^2+\kappa_3 r_*^2\log\frac{1}{\delta} \\
&\max\left(\mathbb{E}[R(\widehat{f})-R(\bar{f})], \mathbb{E}[R_n(\bar{f})-R_n(\widehat{f})]\right) \\
&\quad\leq\quad \kappa_1 L(\bar{f})+(\kappa_2+\kappa_3)r_*^2
\end{aligned}$$

This allows to conclude that $\exists\kappa_4$ such that

$$\mathrm{Var}\left[n(R_n(\bar{f})-R_n(\widehat{f}))\right]\leq n\kappa_4\left(\omega^2\left(r_*\right)+\omega^2\left(\sqrt{L(\bar{f})}\right)\right)$$

If $\omega^2(r)=r/\beta$ (when experiencing random classification noise as in [9]) this implies that the upper-bounds on the expectation and the variance of the empirical excess risk are of the same order of magnitude. Moreover, precise computations for Vapnik-Chervonenkis classes reveal that those upper bounds essentially depend on the model dimension while slowly varying with the sample size.

Those upper-bounds on variance can be completed by upper-bounds on higher moments that suggest that the tails of the empirical excess risk are not larger than the tails of a Gamma distribution. The empirical excess risk satisfies Bernstein inequalities.

Thanks to recent moment inequalities for general functions of independent random variables derived in [4], it is possible to relate the higher moments of the empirical excess risk with the higher moments of the Efron-Stein estimates of

variance. For $q \geq 2$, this observation leads to:

$$\begin{aligned}
&\left\|(Z - \mathbb{E}[Z])_+\right\|_q \\
&\quad \leq \quad \sqrt{3q} \left\| \sqrt{2n \left(P_n(\bar{h} - \widehat{h}_n)^2 + P(\bar{h} - \widehat{h}_n)^2 \right)} \right\|_q \\
&\quad \leq \quad \sqrt{6nq} \left(\sqrt{\left\| P_n(\bar{h} - \widehat{h}_n)^2 \right\|_{q/2}} + \sqrt{\left\| P(\bar{h} - \widehat{h}_n)^2 \right\|_{q/2}} \right) .
\end{aligned}$$

On the other hand, it is possible to check that there exists universal constants κ_5 and κ_6 such that for $q \geq 2$

$$\begin{aligned}
&\left\| P\left(\widehat{h} - \bar{h}\right)^2 \right\|_q \vee \left\| P_n\left(\widehat{h} - \bar{h}\right)^2 \right\|_q \\
&\quad \leq \quad \kappa_5 \left(\omega^2 \left(\sqrt{L(\bar{f})} \right) + \omega^2 \left(r_* \right) \right) + \kappa_6 \omega^2(r_*) q \, .
\end{aligned}$$

Taking advantage on those moment estimates, it is then possible to show that the unnormalized empirical excess risk satisfies a Bernstein-like inequality where the variance term is of the order of $\omega^2(r_*)$. Let $Z = nP_n\left(\bar{h} - \widehat{h}_n\right)$, for $q \geq 2$.

$$\left\| Z - \mathbb{E}[Z] \right\|_q \leq \sqrt{n\kappa_5'} \left(\omega \left(\sqrt{L(\bar{f})} \right) + \omega(r_*) \right) q^{1/2} + \sqrt{n\kappa_6'} \omega(r_*) q \, .$$

References

[1] S .Arlot and P. Massart. Data-driven calibration of penalties for least-squares regression. To appear in *Journal of Machine Learning Research*. 2008.

[2] L. Birgé and P. Massart. Minimal penalties for Gaussian model selection. *Probab. Theory and Related Fields*, 138 (1-2), 33–73, 2007.

[3] S. Boucheron, O. Bousquet, and G. Lugosi. Theory of classification: some recent advances. *ESAIM Probability & Statistics*, pages 329–375, 2006.

[4] S. Boucheron, O. Bousquet, G. Lugosi, and P. Massart. Moment inequalities for functions of independent random variables. *Annals of Probability*, 33(2):514–560, 2005.

[5] S. Boucheron and P. Massart. A poor man's Wilks phenomenon. *to be completed.*

[6] E. Giné and V. Koltchinskii. Concentration inequalities and asymptotic results for ratio type empirical processes. *Annals of Probability*, 34(3):1143–1216, 2006.

[7] E. Giné, V. Koltchinskii, and J. Wellner. *Stochastic Inequalities and Applications*, chapter Ratio limit theorems for empirical processes, pages 249–278. Birkhaüser, 2003.

[8] P. Massart. *Ecole d'Eté de Probabilité de Saint-Flour XXXIII*, chapter Concentration inequalities and model selection. LNM. Springer-Verlag, 2003.

[9] P. Massart and E. Nedellec. Risk bounds for classification. *submitted*, 2003.

[10] M. Talagrand. New concentration inequalities in product spaces. *Inventiones Mathematicae*, 126:505–563, 1996.

[11] A. B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Ann. Statist.*, 32(1):135–166, 2004.

[12] S. van de Geer. *Applications of empirical process theory*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2000.

Some Learning Schemes Generated by Scaling

Ding-Xuan Zhou

Some learning schemes are generated by scaling. In this talk we describe two classes of learning algorithm: Parzen windows for multi-class classification generated by scaling in the time domain and least square regularized regression generated by scaling in the frequency domain or eigenspace.

Parzen windows were introduced for density estimation in the form $p_m(x) = \frac{1}{m\sigma^n}\sum_{i=1}^m \varphi\left(\frac{x-x_i}{\sigma}\right)$ where φ is a density function on $\mathbb{R}^n$ and $\{x_i\}_{i=1}^m$ is a sample drawn from a probability measure ρ_X on the input space $X \subseteq \mathbb{R}^n$. The convergence of $\{p_m(x)\}$ for points in the interior of X has been well studied. We investigate Parzen windows for the purpose of multi-class classification.

Let $Y = \{e_1, \ldots, e_k\}$ be the canonical basis of $\mathbb{R}^k$ representing k classes ($k \geq 2$). Let ρ be a probability measure on $X \times Y$. The misclassification error for a classifier $\mathcal{C} : X \to Y$ is $\mathcal{R}(\mathcal{C}) = \text{Prob}_{(x,y)\in Z}\{\mathcal{C}(x) \neq y\}$.

If we denote $p^j(x) = P(y = e_j|x)$ and $p(x) = (p^1(x), \ldots, p^k(x)) : X \to \mathbb{R}^k$, we see that the best classifier (Bayes rule) f_c minimizing the misclassification error is $f_c(x) = \mathcal{S}(p(x))$ where $\mathcal{S} : \mathbb{R}^k \to Y$ is the splitting function defined as

$$\mathcal{S}(v) = e_{j_v} \text{ where } j_v = \arg\max_{1\leq j\leq k} v^j \quad \text{for} \quad v = (v^1, \ldots, v^k) \in \mathbb{R}^k.$$

The learning ability of a classifier $\mathcal{C}$ is measured by the excess misclassification error $\mathcal{R}(\mathcal{C}) - \mathcal{R}(f_c)$. If a learning algorithm produces a continuous function $f : X \to \mathbb{R}^k$ which induces a classifier $\mathcal{S}(f) : X \to Y$, we can estimate $\mathcal{R}(\mathcal{S}(f)) - \mathcal{R}(f_c)$ by bounding the error in $L^1(X)$ or $C(X)$ as follows [1] where ρ_X is the marginal distribution of ρ on X. This result extends comparison theorems of Zhang (for $k = 2$) and Tewari and Bartlett.

Theorem 1. *For any measurable function $f : X \to \mathbb{R}^k$, we have*

$$\mathcal{R}(\mathcal{S}(f)) - \mathcal{R}(f_c) \leq \sum_{j=1}^k \left\| f^j - p^j \frac{d\rho_X}{dx} \right\|_{L^1(X)}$$

and when f is continuous, with $|X_\rho|$ being the measure of the support of ρ_X,

$$\mathcal{R}(\mathcal{S}(f)) - \mathcal{R}(f_c) \leq 2|X_\rho| \max_{j=1,\ldots,k} \left\| f^j - p^j \frac{d\rho_X}{dx} \right\|_{C(X)}.$$

The Parzen window of order $J \in \mathbb{N}$ is defined by

$$f_{\mathbf{z},\sigma}(x) = \frac{1}{m}\sum_{i=1}^m y_i \Phi\left(\frac{x}{\sigma}, \frac{x_i}{\sigma}\right),$$

where $\sigma > 0$ is a window width, $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m$ is a sample drawn from ρ and $\Phi : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is a basic window function of order J satisfying

(i) $\int_{\mathbb{R}^n} \Phi(x,u)(u-x)^\alpha\, du \equiv \delta_{\alpha,0}$ for $0 < |\alpha| < J$,
(ii) for some $q > n + J$ and $c_q > 0$,

$$|\Phi(x,u)| \le \frac{c_q}{(1+|x-u|)^q} \qquad \forall x, u \in \mathbb{R}^n.$$

The multi-class classifier generated by the Parzen windows is given by $\mathcal{S}(f_{\mathbf{z},\sigma}) : X \to Y$. Its learning ability is studied with methods from approximation theory by considering the behavior of ρ_X near the boundary measured by a Tsybakov type noise condition: with $0 \le \theta \le \infty$ and $C_\theta > 0$,

$$\rho_X\left(\{x \in X : \inf_{y \in \mathbb{R}^n \setminus X} |x-y| \le C_\theta t\}\right) \le t^\theta \qquad \forall t > 0.$$

Theorem 2. *Assume the above noise condition with some $\theta > 0$. If $\frac{d\rho_X}{dx}$ and p^j are Lipschitz s for some $0 < s \le 1$, then with confidence $1-\delta$, we have*

$$\mathcal{R}(\mathcal{S}(f_{\mathbf{z},\sigma})) - \mathcal{R}(f_c) \le \widetilde{C} k \log(2/\delta) m^{-\frac{\beta}{2n+2\beta}},$$

where $\sigma = m^{-\frac{1}{2n+2\beta}}$, $\beta := \min\{s, \frac{\theta(q-n)}{\theta+q-n}\}$ and $\widetilde{C}$ is independent of m, δ or k.

Least square regularized regression is a classical learning algorithm. Here we give an approximation theory viewpoint by considering scaling in the frequency domain or eigenspace. The algorithm is a regularization scheme in a reproducing kernel Hilbert space $\mathcal{H}_K$ associated with a Mercer kernel K on X and with a parameter $\lambda > 0$ takes the form

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

By means of the sampling operator $S_{\mathbf{x}} : \mathcal{H}_K \to \mathbb{R}^m$ defined as $S_{\mathbf{x}}(f) = \big(f(x_i)\big)_{i=1}^m$ and its adjoint $S_{\mathbf{x}}^T$, it can be represented as $f_{\mathbf{z},\lambda} = \big(\frac{1}{m} S_{\mathbf{x}}^T S_{\mathbf{x}} + \lambda I\big)^{-1} \frac{1}{m} S_{\mathbf{x}}^T \mathbf{y}$ where $\mathbf{y} = (y_i)_{i=1}^m \in \mathbb{R}^m$. The operator $\frac{1}{m} S_{\mathbf{x}}^T S_{\mathbf{x}}$ is a good approximation of the integral operator L_K on $\mathcal{H}_K$ or $L^2_{\rho_X}$ given by $L_K(f) = \int_X K_v f(v) d\rho_X$. Hence

$$f_{\mathbf{z},\lambda} = \frac{1}{m} \sum_{i=1}^m y_i \big(\frac{1}{m} S_{\mathbf{x}}^T S_{\mathbf{x}} + \lambda I\big)^{-1} (K_{x_i}) \approx \frac{1}{m} \sum_{i=1}^m y_i \Phi_\lambda(\cdot, x_i),$$

where $\Phi_\lambda(\cdot, x) = \big(L_K + \lambda I\big)^{-1}(K_x)$. But $y_i \approx f_\rho(x_i)$ where f_ρ is the regression function, so $f_{\mathbf{z},\lambda} \approx \frac{1}{m} \sum_{i=1}^m f_\rho(x_i) \Phi_\lambda(\cdot, x_i) \approx \int_X \Phi_\lambda(\cdot, x) f_\rho(x) d\rho_X$.

Choose a normalized eigenpairs of L_K in $L^2_{\rho_X}$ as $\{(\lambda_i, \varphi_i)\}_{i \ge 1}$. By the Mercer Theorem, we have $K_x = \sum_{i \ge 1} \lambda_i \varphi_i(x) \varphi_i$. Hence $\Phi_\lambda(u, x) = \sum_{i \ge 1} \frac{\lambda_i}{\lambda_i + \lambda} \varphi_i(x) \varphi_i(u)$. Since the delta function can be expressed as $\delta_x(u) = \sum_{i \ge 1} \varphi_i(x) \varphi_i(u)$, we see that the kernel function Φ_λ is a scaled version of the delta function by scaling 1 to $\frac{\lambda_i}{\lambda_i + \lambda} \to 1$ (as $\lambda \to 0$). More quantitatively, we have the following bound [2].

Theorem 3. *Let $\frac{1}{2} < r \le \frac{3}{2}$ and $L_K^r(L^2_{\rho_X})$ be the range of L_K on $L^2_{\rho_X}$. Then*

$$\left\| \int_X \Phi_\lambda(u, x) \cdot (x) d\rho_X(x) - I \right\|_{L_K^r(L^2_{\rho_X}) \to \mathcal{H}_K} \le \lambda^{r - \frac{1}{2}}.$$

References

[1] Z. W. Pan, D. H. Xiang, Q. W. Xiao, and D. X. Zhou, *Parzen windows for multi-class classification*, J. Complexity, to appear.

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

Nonparametric Regression between Manifolds

Matthias Hein

(joint work with Florian Steinke, Bernhard Schölkopf)

We consider the problem of learning a mapping $\phi : M \to N$, where M, N are Riemannian manifolds, given k i.i.d. samples $(X_i, Y_i)_{i=1}^k$, $X_i \in M$ and $Y_i \in N$, from a probability measure P on $M \times N$. This learning problem reduces to standard multivariate regression if M and N are both Euclidean spaces $\mathbb{R}^m$ and $\mathbb{R}^n$ and to regression on a manifold if at least N is Euclidean. For the case, where M is Euclidean and N is a manifold, first results have been obtained in [1]. The general setting together with a more formal approach is presented in [2].

We use the squared geodesic distance as loss measure and solve the problem using regularized empirical risk minimization, which can be formulated in our setting as

$$\operatorname*{argmin}_{\phi \in \mathcal{F}} \frac{1}{k} \sum_{i=1}^{k} d_N^2(Y_i, \phi(X_i)) + \lambda\, S(\phi), \tag{1}$$

where $\mathcal{F} \subset C^2(M, N)$, d_N is the metric on N, $\lambda \in \mathbb{R}_+$ the regularization parameter, and $S : C^2(M, N) \to \mathbb{R}$ the regularization functional. The goal is to find the Bayes optimal mapping $\eta : M \to N$ (the regression function), which is defined as

$$\eta = \operatorname*{argmin}_{\phi : M \to N,\ \phi \text{measurable}} \mathbb{E}[d_N^2(\phi(X), Y)].$$

The purpose of the regularizer $S(\phi)$ is to avoid overfitting of the training data. It is known that first order regularizers lead generally to piecewise geodesic solutions [3], which is usually not sufficiently smooth for applications. We propose to use a second order regularizer, which we have named Eells energy in honor of James Eells, one of the pioneers of harmonic mappings,

$$S_{\text{Eells}}(\phi) = \int_M \|\nabla' d\phi\|^2_{T_x^* M \otimes T_x^* M \otimes T_{\phi(x)} N} dV(x),$$

where ∇' is the pull-back connection. The Eells energy reduces to the thin-plate-spline energy, in the case where M and N are Euclidean.

The null space of a regularizer is quite important, since it contains the mappings which are not penalized that is the set of mappings which ones is interested to fit the data with. Interestingly, the null space of the Eells energy are the so called totally geodesic maps, which can be seen as the generalization of linear mappings in Euclidean space to mappings between Riemannian manifolds.

Given that the input and output manifold are isometrically embedded into Euclidean space one can find a (local) minima of the problem (1) quite efficiently as shown in [1].

It turns out that the problem of nonparametric regression between manifolds is a non-trivial generalization of the standard regression problem in Euclidean space. We illustrate this non-trivial structure with two open problems.

Homotopy classes: The first one is the non-trivial topological structure of the set $C(M, N)$, the continuous mappings from M to N. Namely, the homotopy classes $[M, N]$ (the equivalence classes of mappings which can be continuously transformed into each other) can be non-trivial. This has an important practical effect, since the objective function in (1) is usually optimized using descent techniques. However, this implies that the initial mapping for the gradient descent fixes the homotopy class, which leads to the question:

Does there exist a way to construct an initial solution $\phi^{(0)}$ which for sufficiently large sample size k is guaranteed to lie in the same homotopy class as η ?

We have found a first positive result for mappings from S^1 to S^1.

Capacity of totally geodesic mappings: Linear mappings in Euclidean space are known to have small capacity. It turns out that totally geodesic mappings can even have infinite capacity as the following reformulation of a well-known result in number theory, see [4], shows.

Theorem 1. *Let $(X_i, Y_i) \in S^1 \times S^1$, $i = 1, \ldots, k$, be the training data. Then there exists for any set of training data and any $\epsilon > 0$ a $K \in \mathbb{N}$ such that*

$$\max_{i=1,\ldots,k} d_N(\phi_K(X_i), Y_i) \leq \epsilon,$$

where $\phi_K : S^1 \to S^1$, $\phi_K(x) = \mathrm{mod}(Kx + \delta, 2\pi)$.

This leads us to the next open problem:

Under which conditions on M and N has the set of totally geodesic mappings finite capacity ?

References

[1] F. Steinke, M. Hein, J. Peters and B. Schölkopf, *Manifold-valued Thin-Plate Splines with Applications in Computer Graphics*, Computer Graphics Forum, **27** (2008), 437-448.

[2] M. Hein, F. Steinke and B. Schölkopf, *Energy functionals for manifold-valued mappings and their properties*, Technical Report 167, Max Planck Institute for Biological Cybernetics, January 2008.

[3] F. Memoli, G. Sapiro and S. Osher, *Solving variational problems and partial differential equations mapping into general target manifolds*, J. Comp. Phys., **195** (2004), 263-292.

[4] T. M. Apostol, *Modular Functions and Dirichlet Series in Number Theory*, Springer, New York, (1990).

Pricing of American Options by Regression-based Monte Carlo Methods

Michael Kohler
(joint work with Daniel Egloff, Adam Krzyżak, Nebojsa Todorovic)

In this article we consider American options in discrete time, which are also called Bermuda options. The price V_0 of such options can be defined as a solution of an optimal stopping problem

$$V_0 = \sup_{\tau \in T(0,\ldots,T)} E\{f_\tau(X_\tau)\}. \tag{1}$$

Here f_t is the (discounted) payoff function, X_0, X_1, ..., X_T is the underlying stochastic process describing e.g. the prices of the underlyings and the financial environment (like interest rates, etc.) which we assume to be a R^d-valued Markov process, and $T(0,\ldots,T)$ is the class of all $\{0,\ldots,T\}$-valued stopping times, i.e., $\tau \in T(0,\ldots,T)$ is a measurable function of $X_0, \ldots, X_T$ satisfying

$$\{\tau = \alpha\} \in F(X_0,\ldots,X_\alpha) \quad \text{for all } \alpha \in \{0,\ldots,T\}.$$

The computation of (1) can be done by determination of an optimal stopping rule $\tau^* \in T(0,\ldots,T)$ satisfying

$$V_0 = E\{f_{\tau^*}(X_{\tau^*})\}. \tag{2}$$

Let

$$q_t(x) = \sup_{\tau \in T(t+1,\ldots,T)} E\{f_\tau(X_\tau)|X_t = x\} \tag{3}$$

be the so–called continuation value describing the value of the option at time t given $X_t = x$ and subject to the constraint of holding the option at time t rather than exercising it. Here $T(t+1,\ldots,T)$ is the class of all $\{t+1,\ldots,T\}$–valued stopping times. It can be shown that

$$\tau^* = \inf\{s \geq 0 : q_s(X_s) \leq f_s(X_s)\} \tag{4}$$

satisfies (2), i.e., τ^* is an optimal stopping time. Therefore it suffices to compute the continuation values (3) in order to solve the optimal stopping problem (1).

It is easy to see that the continuation values satisfy the dynamic programming equations

$$\begin{aligned} q_T(x) &= 0, \\ q_t(x) &= E\{\max\{f_{t+1}(X_{t+1}), q_{t+1}(X_{t+1})\}|X_t = x\} \\ &\quad (t = 0,1,\ldots,T-1). \end{aligned} \tag{5}$$

Unfortunately, the conditional expectation in (5) in general cannot be computed in applications. The basic idea of regression-based Monte Carlo methods for pricing American options is to apply recursively regression estimates to artificially created samples of

$$(X_t, \max\{f_{t+1}(X_{t+1}), \hat{q}_{t+1}(X_{t+1})\})$$

(so–called Monte Carlo samples) to construct estimates $\hat{q}_t$ of q_t. In connection with linear regression this was proposed in Tsitsiklis and Van Roy (1999), and, based on a different regression estimation than (5), in Longstaff and Schwartz (2001).

In this article we propose to use various nonparametric regression estimates in order to compute the conditional expectations in (5). The outline of the estimation procedure is as follows:

(1) Set $q_{n,T} = 0$.

(2) Recursively define estimates of q_{T-1}, q_{T-2}, $\ldots$, q_0:

(a) Create Monte Carlo samples $\{X_{i,s}^{(t)}\}_{s=0,\ldots,T}$ of size n of $\{X_s\}_{s=0,\ldots,T}$.

(b) Use

$$X_{i,t+1}^{(t)}, \ldots, X_{i,T}^{(t)}, q_{n,t+1}, \ldots, q_{n,T} \quad (i = 1, \ldots, n)$$

in order to compute $\hat{Y}_{i,t}^{(t)} \quad (i = 1, \ldots, n)$.

(c) Use

$$\left\{ (X_{i,t}^{(t)}, \hat{Y}_{i,t}^{(t)}) \right\}_{i=1,\ldots,n}$$

to construct a (fully data-driven) regression estimate $q_{n,t}$ of q_t.

(3) Estimate

$$\tau^* = \inf \{ s \geq 0 \, : \, q_s(X_s) \leq f_s(X_s) \}$$

by

$$\hat{\tau} = \inf \{ s \geq 0 \, : \, q_{n,s}(X_s) \leq f_s(X_s) \},$$

and

$$V_0 = E \left[f_{\tau^*}(X_{\tau^*}) \right]$$

by the corresponding Monte Carlo estimate.

Estimates are defined as above by using truncated versions of least squares splines (cf., e.g., Chapter 15 in Györfi et al. (2002)) for estimate $q_{n,t}^{(1)}$, least squares neural networks (cf., e.g., Chapter 16 in Györfi et al. (2002)) for estimate $q_{n,t}^{(2)}$ and smoothing splines (cf., e.g., Chapter 20 in Györfi et al. (2002)) for estimate $q_{n,t}^{(3)}$, where the parameter of the estimate is chosen by splitting of the sample.

The following results are valid for bounded Markov processes and bounded payoff functions.

Theorem 1. (Consistency)

a) *(Egloff, Kohler and Todorovic (2007)).*

$$\int |q_{n,t}^{(1)}(x) - q_t(x)|^2 P_{X_t}(dx) \to 0 \quad \textit{in probability for all } t \in \{0, \ldots, T-1\}.$$

b) *(Kohler, Krzyżak and Todorovic (2006)).*

$$\int |q_{n,t}^{(2)}(x) - q_t(x)|^2 P_{X_t}(dx) \to 0 \quad \textit{in probability for all } t \in \{0, \ldots, T-1\}.$$

c) *(Kohler (2008)).*

$$\int |q_{n,t}^{(3)}(x) - q_t(x)|^2 P_{X_t}(dx) \to 0 \quad \textit{in probability for all } t \in \{0, \dots, T-1\}.$$

Theorem 2. (Rate of convergence)

a) *(Egloff, Kohler and Todorovic (2007)). Assume that the continuation values are (p, C)–smooth, i.e., that derivatives of order $p-1$ exists and are Lipschitz continuous with Lipschitz constant C. Then for all $t \in \{0, 1, \dots, T-1\}$:*

$$\int |q_{n,t}^{(1)}(x) - q_t(x)|^2 P_{X_t}(dx) = O_P\left(C^{2d/(2p+d)} \cdot \left(\frac{\log n}{n}\right)^{2p/(2p+d)}\right)$$

b) *(Kohler (2008)). Assume*

$$J_k(q_t) = \sum_{\alpha_1,\dots,\alpha_d \in N,\, \alpha_1+\dots+\alpha_d = k} \frac{k!}{\alpha_1! \cdot \dots \cdot \alpha_d!} \int_{R^d} \left| \frac{\partial^k q_t}{\partial x_1^{\alpha_1} \dots \partial x_d^{\alpha_d}}(x) \right|^2 dx \le C$$

for all $t \in \{0, \dots, T-1\}$ and some natural number k. Then for all $t \in \{0, \dots, T-1\}$:

$$\int |q_{n,t}^{(3)}(x) - q_t(x)|^2 P_{X_t}(dx) = O_P\left(C^{2d/(2k+d)} \cdot \left(\frac{\log n}{n}\right)^{2k/(2k+d)}\right)$$

c) *(Kohler, Krzyżak and Todorovic (2006)).*
Assume that the Fourier transform $\hat{F}_{q_t}(\omega) = \frac{1}{(2\pi)^{d/2}} \int_{R^d} e^{-i \cdot \omega^T z} q_t(z) dz$ $(\omega \in R^d)$ of q_t satisfies

$$q_t(x) = \frac{1}{(2\pi)^{d/2}} \int_{R^d} e^{i \cdot \omega^T x} \hat{F}_{q_t}(\omega) d\omega \quad (x \in R^d)$$

and

$$\int_{R^d} \|\omega\| \cdot \hat{F}_{q_t}(\omega) d\omega < \infty$$

for all $t \in \{0, \dots, T-1\}$. Then for all $t \in \{0, 1, \dots, T-1\}$:

$$\int |q_{n,t}^{(2)}(x) - q_t(x)|^2 P_{X_t}(dx) = O_P\left(\sqrt{\frac{\log^5 n}{n}}\right)$$

References

[1] D. Egloff, M. Kohler and N. Todorovic, *A dynamic look-ahead Monte Carlo algorithm for pricing American options*, Annals of Applied Probability **17** (2007), 1138–1171.

[2] L. Györfi, M. Kohler, A. Krzyżak and H. Walk, *A Distribution-Free Theory of Nonparametric Regression.* Springer Series in Statistics, Springer (2002).

[3] M. Kohler, *A regression based smoothing spline Monte Carlo algorithm for pricing American options*, AStA Advances in Statistical Analysis **92**, (2008), 153–178.

[4] M. Kohler, A. Krzyżak and N. Todorovic, *Pricing of high-dimensional American options by neural networks*, submitted for publication (2006).

[5] F. A. Longstaff and E. S. Schwartz, *Valuing American options by simulation: a simple least-squares approach*, Review of Financial Studies **14** (2001), 113–147.

[6] J. N. Tsitsiklis and B. Van Roy, *Optimal stopping of Markov processes: Hilbert space theory, approximation algorithms, and an application to pricing high-dimensional financial derivatives*, IEEE Trans Autom. Control **44** (1999), 1840–1851.

Approximation and Balancing Properties of Wavelet Frames

Bin Han

In the interesting work [7, 8], Jetter and Zhou studied the approximation property of a family of quasi-interpolation operators. More precisely, they prove that

Jetter-Zhou [7, 8]**:** *Let $\varphi \in L_2(\mathbb{R})$ and $\nu \geq 0$. Define a linear operator P by $P(f) := \sum_{k\in\mathbb{Z}} \langle f, \varphi(\cdot - k)\rangle \varphi(\cdot - k)$, $f \in L_2(\mathbb{R})$. Then $\|f - P(f)\|_{L_2(\mathbb{R})} \leq C_\varphi |f|_{H^\nu(\mathbb{R})}$ for all $f \in H^\nu(\mathbb{R})$ with a positive constant*

$$C_\varphi := \pi^{-1/2} \sqrt{\max(c_1, c_3) + \max(2c_2, 2c_4 + 1)},$$

provided that there exist positive constants c_1, c_2, c_3, c_4 such that for almost every $\xi \in [-\pi, \pi]$, the following inequalities hold

$$\big|1 - |\hat{\varphi}(\xi)|^2\big|^2 \leq c_1 |\xi|^{2\nu}, \quad \sum_{k\in\mathbb{Z}\setminus\{0\}} |\hat{\varphi}(\xi)|^2 |\hat{\varphi}(\xi + 2\pi k)|^2 \leq c_2 |\xi|^{2\nu},$$

$$\sum_{k\in\mathbb{Z}\setminus\{0\}} |\xi + 2\pi k|^{-2\nu} |\hat{\varphi}(\xi)|^2 |\hat{\varphi}(\xi + 2\pi k)|^2 \leq c_3,$$

$$\sum_{k\in\mathbb{Z}\setminus\{0\}} |\xi + 2\pi k|^{-2\nu} \sum_{\ell\in\mathbb{Z}\setminus\{0\}} |\hat{\varphi}(\xi + 2\pi\ell)|^2 |\hat{\varphi}(\xi + 2\pi k)|^2 \leq c_4.$$

In this talk, we shall discuss several applications of the above result to study some frame approximation properties of wavelet frames, in particular, stationary tight wavelet frames, pairs of dual multiframelets, nonstationary tight wavelet frames, and pairs of dual wavelet frames in Sobolev spaces.

For stationary tight wavelet frames obtained via the following Oblique Extension Principle (OEP) in [2] and independently [1], we have the following result.

Daubechies-Han-Ron-Shen [2]**:** *Let $\phi \in L_2(\mathbb{R})$ be a compactly supported function satisfying $\hat{\phi}(2\xi) = \hat{a}(\xi)\hat{\phi}(\xi)$. Suppose that $\hat{a}, \hat{\Theta}, \widehat{b^1}, \ldots, \widehat{b^L}$ are 2π-periodic trigonometric polynomials satisfying $\hat{\Theta}(0) = \hat{a}(0) = 1$ and*

$$\hat{\Theta}(2\xi)|\hat{a}(\xi)|^2 + \sum_{\ell=1}^{L} |\widehat{b^\ell}(\xi)|^2 = \hat{\Theta}(\xi), \quad \hat{\Theta}(2\xi)\hat{a}(\xi)\overline{\hat{a}(\xi + \pi)} + \sum_{\ell=1}^{L} \widehat{b^\ell}(\xi)\overline{\widehat{b^\ell}(\xi + \pi)} = 0.$$

Define $\widehat{\psi^\ell}(2\cdot) := \widehat{b^\ell}\hat{\phi}$. Then $X(\psi^1, \ldots, \psi^L) := \{\psi^\ell_{j,k} := 2^{j/2}\psi^\ell(2^j \cdot - k) : \ell = 1, \ldots, L, j, k \in \mathbb{Z}\}$ is a tight wavelet frame in L_2: $\|f\|^2_{L_2} = \sum_{\ell=1}^{L} \sum_{j,k\in\mathbb{Z}} |\langle f, \psi^\ell_{j,k}\rangle|^2$ for $f \in L_2(\mathbb{R})$. Moreover, if $\hat{a}$ contains the factor $(1 + e^{-i\xi})^m$, then it has the frame

approximation order ν with $\nu \le m$:

$$\|f-Q_n(f)\|_{L_2(\mathbb{R})} \le C|f|_{H^\nu(\mathbb{R})},\ n\in\mathbb{N} \quad \textit{with} \quad Q_n(f) := \sum_{\ell=1}^{L}\sum_{j=-\infty}^{n-1}\sum_{k\in\mathbb{Z}}\langle f,\psi_{j,k}^\ell\rangle\psi_{j,k}^\ell,$$

if and only if $\hat\Theta(\xi)-\hat\Theta(2\xi)|\hat a(\xi)|^2 = O(|\xi|^\nu),\ \xi\to 0$.

For nonstationary tight wavelet frames, we have the following result.

Han-Shen [4]**:** *Let $\{\widehat{a_j}\}_{j=1}^\infty$ be 2π-periodic trigonometric polynomials such that $\widehat{a_j}(0)=1$ and $\sum_{j=1}^\infty 2^{-j}\deg(\widehat{a_j})<\infty$. Suppose that there are 2π-periodic trigonometric polynomials $\widehat{b_j^\ell}$, $\ell=1,\dots,\mathcal{J}_j$, satisfying*

$$|\widehat{a_j}(\xi)|^2+\sum_{\ell=1}^{\mathcal{J}_j}|\widehat{b_j^\ell}(\xi)|^2=1,\quad \widehat{a_j}(\xi)\overline{\widehat{a_j}(\xi+\pi)}+\sum_{\ell=1}^{\mathcal{J}_j}\widehat{b_j^\ell}(\xi)\overline{\widehat{b_j^\ell}(\xi+\pi)}=0.$$

Define nonstationary refinable functions ϕ_{j-1} and wavelets ψ_{j-1}^ℓ by $\widehat{\phi_{j-1}}(\xi) := \prod_{n=1}^\infty \widehat{a_{n+j-1}}(2^{-n}\xi)$ and $\widehat{\psi_{j-1}^\ell} := \widehat{b_j^\ell}(\cdot/2)\widehat{\phi_j}(\cdot/2)$. Denote $\psi_{j;j,k}^\ell := 2^{j/2}\psi^\ell(2^j\cdot-k)$. Then $X(\phi_0;\{\psi_j^\ell\}_{j\in\mathbb{N}_0,\ell\in\{1,\dots,\mathcal{J}_{j+1}\}})$ is a nonstationary tight wavelet frame in $L_2(\mathbb{R})$:

$$\|f\|_{L_2(\mathbb{R})}^2=\sum_{k\in\mathbb{Z}}|\langle f,\phi_0(\cdot-k)\rangle|^2+\sum_{j=0}^\infty\sum_{\ell=1}^{\mathcal{J}_{j+1}}\sum_{k\in\mathbb{Z}}|\langle f,\psi_{j;j,k}^\ell\rangle|^2,\qquad f\in L_2(\mathbb{R}).$$

If we further assume that for some $\alpha\ge 0$ and $0\le\beta<1$, $\deg(\widehat{a_j})=O(j^\alpha 2^{\beta j})$ as $j\to\infty$ and there exist $\nu\in\frac12\mathbb{N}$ and $N\in\mathbb{N}$ such that for $j\ge N$, $|\widehat{a_j}(\xi)|^2 = 1+O(|\xi|^{2\nu})$, $\xi\to 0$, then we have the weak frame approximation property:

$$\|f-Q_n(f)\|_{L_2(\mathbb{R})}\le Cn^{\nu\alpha}2^{-\nu(1-\beta)n}|f|_{H^\nu(\mathbb{R})}\ \forall\, f\in H^\nu(\mathbb{R}),\ n\ge N,$$

where $Q_n(f) := \sum_{k\in\mathbb{Z}}\langle f,\phi_0(\cdot-k)\rangle\phi_0(\cdot-k)+\sum_{j=0}^{n-1}\sum_{\ell=1}^{\mathcal{J}_{j+1}}\sum_{k\in\mathbb{Z}}\langle f,\psi_{j;j,k}^\ell\rangle\psi_{j;j,k}^\ell$.

Moreover, the nonstationary tight wavelet frame $X(\phi_0;\{\psi_j^\ell\}_{j\in\mathbb{N}_0,\ell=1,\dots,\mathcal{J}_j})$ constructed above provides the frame approximation order $\nu>0$:

$$\|f-Q_n(f)\|_{L_2(\mathbb{R})}\le C2^{-\nu n}|f|_{H^\nu(\mathbb{R})}\quad\forall\, f\in H^\nu(\mathbb{R}),\quad \text{large } n$$

if and only if there is $C>0$ such that for all large n, for almost every $\xi\in[-\pi,\pi]$,

$$\left|1-|\widehat{\phi_n}(\xi)|^2\right|^2\le C|\xi|^{2\nu},\qquad \sum_{k\in\mathbb{Z}\setminus\{0\}}|\widehat{\phi_n}(\xi)|^2|\widehat{\phi_n}(\xi+2\pi k)|^2\le C|\xi|^{2\nu}.$$

In particular, the two conditions hold if $1-|\widehat{\phi_n}(\xi)|^2\le C|\xi|^{2\nu}$, *a.e.* $\xi\in[-\pi,\pi]$.

If we choose $\widehat{a_j}(\xi) := \cos^{2j}(\xi/2)\sum_{j=0}^{l_j-1}\binom{m+j-1}{j}\sin^{2j}(\xi/2)$ with $\liminf_{j\to\infty}\frac{l_j}{j}>0$, then all ϕ_j,ψ_j^ℓ are $C^\infty(\mathbb{R})$ functions with symmetry and the nonstationary tight

wavelet frame has the frame approximation order ν for all $\nu > 0$. By [6], for any $s \in \mathbb{R}$, there exist positive constants C_1 and C_2 such that

$$C_1\|f\|^2_{H^s(\mathbb{R})} \le \sum_{k\in\mathbb{Z}} |\langle f, \phi_0(\cdot - k)\rangle|^2 + \sum_{j=0}^{\infty}\sum_{\ell=1}^{\mathcal{J}_{j+1}}\sum_{k\in\mathbb{Z}} 2^{2js}|\langle f, \psi^\ell_{j;j,k}\rangle|^2 \le C_2\|f\|^2_{H^s(\mathbb{R})}.$$

Namely, $\{\phi_0(\cdot - k) : k \in \mathbb{Z}\} \cup \{2^{j/2-s}\psi^\ell_j(2^j \cdot -k) : j \in \mathbb{N}_0, k \in \mathbb{Z}, \ell = 1, \dots, \mathcal{J}_{j+1}$ is a wavelet frame in the Sobolev space $H^\nu(\mathbb{R})$ for all $\nu \in \mathbb{R}$.

For pairs of dual multiframelets, we have the following result.

Han [3]**:** *Let ϕ and $\tilde\phi$ be two $r\times 1$ refinable function vectors of compactly supported functions in $L_2(\mathbb{R})$ satisfying $\hat\phi(2\xi) = \hat a(\xi)\hat\phi(\xi)$ and $\hat{\tilde\phi}(2\xi) = \hat{\tilde a}(\xi)\hat{\tilde\phi}(\xi)$. Suppose $\hat a$ and $\hat{\tilde a}$ have m and $\tilde m$ sum rules. If $r > 1$, then one can obtain in a constructive way $r \times r$ matrices $\hat\Theta, \hat{\tilde\Theta}$, $\widehat{b^\ell}, \widehat{\tilde b^\ell}$, $\ell = 1, \dots, d$ of trigonometric polynomials such that all the conditions in OEP are satisfied with $L = 2$, and*

(i) *$\hat\Theta^{-1}$ and $\hat{\tilde\Theta}^{-1}$ are $r \times r$ matrices of 2π-periodic trigonometric polynomials.*
(ii) *$(X(\{\psi^1, \psi^2\}), X(\{\tilde\psi^1, \tilde\psi^2\}))$ is a pair of compactly supported dual wavelet frames in $L_2(\mathbb{R})$ with $\tilde m$ and m vanishing moments, respectively.*
(iii) *No de-convolution is involved in its fast frame transform with the highest possible balancing order m and the frame approximation order m.*

Han [3]**:** *Let ϕ and $\tilde\phi$ be two compactly supported scalar spline refinable functions in $L_2(\mathbb{R})$ with masks a and $\tilde a$. For any 2π-periodic trigonometric polynomials $\hat\Theta, \hat{\tilde\Theta}$, $\widehat{b^\ell}, \widehat{\tilde b^\ell}$, $\ell = 1, \dots, L$, such that OEP holds, if no de-convolution is involved in its fast frame transform, then the pair of dual d-wavelet frames obtained via OEP with $r = 1$ can have the frame approximation order at most two and vanishing moments at most one.*

Finally, based on [5], we discuss pairs of dual wavelet frames in a pair of Sobolev spaces. For $\phi, \psi^1, \dots, \psi^L \in H^s(\mathbb{R}^d)$, we say that $X^s(\phi; \psi^1, \dots, \psi^L) :=$

$$\{\phi(\cdot - k) : k \in \mathbb{Z}^d\} \cup \big\{\psi^{\ell,s}_{j,k} := 2^{j(d/2-s)}\psi^\ell(2^j\cdot - k) : j \in \mathbb{N}_0,\ k \in \mathbb{Z}^d, \ell = 1, \dots, L\big\},$$

is a wavelet frame in $H^s(\mathbb{R}^d)$ if there are positive constants C_1 and C_2 such that

$$C_1\|f\|^2_{H^s} \le \sum_{k\in\mathbb{Z}^d} |\langle f, \phi(\cdot - k)\rangle_{H^s}|^2 + \sum_{\ell=1}^{L}\sum_{j=0}^{\infty}\sum_{k\in\mathbb{Z}^d} |\langle f, \psi^{\ell,s}_{j,k}\rangle_{H^s}|^2 \le C_2\|f\|^2_{H^s}, f \in H^s.$$

We say that $(X^s(\phi; \psi^1, \dots, \psi^L), X^{-s}(\tilde\phi; \tilde\psi^1, \dots, \tilde\psi^L))$ is a pair of dual wavelet frames in $(H^s(\mathbb{R}^d), H^{-s}(\mathbb{R}^d))$ if $X^s(\phi; \psi^1, \dots, \psi^L)$ is a wavelet frame in $H^s(\mathbb{R}^d)$, $X^{-s}(\tilde\phi; \tilde\psi^1, \dots, \tilde\psi^L)$ is a wavelet frame in $H^{-s}(\mathbb{R}^d)$, and for $f \in H^s$ and $g \in H^{-s}$, $\langle f, g\rangle = \sum_{k\in\mathbb{Z}^d}\langle f, \tilde\phi(\cdot - k)\rangle\langle\phi(\cdot - k), g\rangle + \sum_{\ell=1}^{L}\sum_{j=0}^{\infty}\sum_{k\in\mathbb{Z}^d}\langle f, \tilde\psi^{\ell,-s}_{j,k}\rangle\langle\psi^{\ell,s}_{j,k}, g\rangle$.

Complete characterizations and examples in [5] are discussed for pairs of dual wavelet frames and dual Riesz wavelets in a pair of Sobolev spaces.

References

[1] C. K. Chui, W. He and J. Stöckler, *Compactly supported tight and sibling frames with maximum vanishing moments*, Appl. Comput. Harmon. Anal. **13** (2002), 224–262.

[2] I. Daubechies, B. Han, A. Ron, and Z. Shen, *Framelets: MRA-based constructions of wavelet frames*, Appl. Comput. Harmon. Anal. **14** (2003), 1–46.

[3] B. Han, *Dual multiwavelet frames with high balancing order and compact fast frame transform*, Appl. Comput. Harmon. Anal., to appear.

[4] B. Han and Z. Shen, *Compactly supported symmetric C^∞ wavelets with spectral approximation order*, SIAM J. Math. Anal., to appear.

[5] B. Han and Z. Shen, *Dual wavelet frames and Riesz bases in Sobolev spaces*, Constr. Approx., to appear.

[6] B. Han and Z. Shen, *Characterization of Sobolev spaces of arbitrary smoothness using nonstationary tight wavelet frames*, Israel J. Math., to appear.

[7] K. Jetter and D. X. Zhou, *Order of linear approximation from shift-invariant spaces.* Constr. Approx. **11** (1995), 423–438.

[8] K. Jetter and D. X. Zhou, *Approximation order of linear operators and finitely generated shift-invariant spaces*, preprint, (1998).

RKHS Representation of Measures Applied to Homogeneity, Independence, and Fourier Optics

Bernhard Schölkopf

(joint work with Bharath Sriperumbudur, Arthur Gretton, Kenji Fukumizu)

A symmetric function $k : \mathcal{X}^2 \to \mathbb{R}$, where $\mathcal{X}$ is a nonempty set, is called a positive definite (pd) kernel if for arbitrary points $x_1, \ldots, x_m \in \mathcal{X}$ and coefficients $a_1, \ldots, a_m \in \mathbb{R}$, we have

$$\sum_{i,j} a_i a_j k(x_i, x_j) \geq 0.$$

The kernel is called strictly positive definite if for pairwise distinct points, the implication $\sum_{i,j} a_i a_j k(x_i, x_j) = 0 \Longrightarrow \forall i : a_i = 0$ is valid.

Any positive definite kernel induces a mapping

$$x \mapsto k(x, .)$$

into a reproducing kernel Hilbert space (RKHS) satisfying

$$\langle k(x, .), k(x', .)\rangle = k(x, x')$$

for all $x, x' \in \mathcal{X}$.

Consider two sets of points $X := \{x_1, \ldots, x_m\} \subset \mathcal{X}, Y := \{y_1, \ldots, y_n\} \subset \mathcal{X}$. We define the mean map μ by

$$\mu(X) = \frac{1}{m} \sum_{i=1}^{m} k(x_i, \cdot).$$

One can define a classification rule in $\mathcal{H}$ based on the closest mean, i.e., using a hyperplane with normal vector $\mu(X) - \mu(Y)$ [4]. This begs the question: when is this normal vector zero (in which case it does not define a hyperplane)? For polynomial kernels $k(x, x') = (\langle x, x'\rangle + 1)^d$, this amounts to all empirical moments

up to order d vanishing. For strictly positive definite kernels, the means coincide only if $X = Y$, rendering μ injective:

Lemma. *Assume X, Y are defined as above, k is strictly pd, and for all i, j, $x_i \neq x_j$, and $y_i \neq y_j$. If for some $\alpha_i, \beta_j \in \mathbb{R} \setminus \{0\}$, we have*

$$\sum_{i=1}^{m} \alpha_i k(x_i, .) = \sum_{j=1}^{n} \beta_j k(y_j, .), \tag{1}$$

then $X = Y$.

To see this, assume w.l.o.g. that $x_1 \notin Y$. Subtract $\sum_{j=1}^{n} \beta_j k(y_j, .)$ from (1), and make it a sum over pairwise distinct points, to get

$$0 = \sum_i \gamma_i k(z_i, .),$$

where $z_1 = x_1, \gamma_1 = \alpha_1 \neq 0$, and $z_2, \cdots \in X \cup Y \setminus \{x_1\}$, $\gamma_2, \cdots \in \mathbb{R}$. Take the RKHS dot product with $\sum_j \gamma_j k(z_j, .)$ to get

$$0 = \sum_{ij} \gamma_i \gamma_j k(z_i, z_j),$$

with $\gamma \neq 0$, hence k cannot be strictly pd. □

The mean map has some other interesting properties. Among them is the fact that $\mu(X)$ represents the operation of taking a mean of a function on the sample X:

$$\langle \mu(X), f \rangle = \left\langle \frac{1}{m} \sum_{i=1}^{m} k(x_i, \cdot), f \right\rangle = \frac{1}{m} \sum_{i=1}^{m} f(x_i)$$

Moreover, we have

$$\|\mu(X) - \mu(Y)\| = \sup_{\|f\| \leq 1} |\langle \mu(X) - \mu(Y), f \rangle| = \sup_{\|f\| \leq 1} \left| \frac{1}{m} \sum_{i=1}^{m} f(x_i) - \frac{1}{n} \sum_{i=1}^{n} f(y_i) \right|.$$

If $\mathbf{E}_{x,x' \sim p}[k(x, x')]$, $\mathbf{E}_{x,x' \sim q}[k(x, x')] < \infty$, then the above statements generalize to Borel measures p, q, with the difference being that the mean map is defined as

$$\mu \colon p \mapsto \mathbf{E}_{x \sim p}[k(x, \cdot)],$$

and the notion of strictly pd kernels is replaced by that of characteristic kernels [1]. In this case, the mean map can be viewed as a generalization of the *moment generating function* M_p of a random variable x with distribution p,

$$M_p(.) = \mathbf{E}_{x \sim p} \left[e^{\langle x, \cdot \rangle} \right].$$

If we restrict the class of distributions, the class of kernels for which μ is injective gets larger. To see this, consider a bounded translation invariant kernel $k(x, x') = \psi(x - x')$, with continuous $\psi : \mathbb{R}^d \to \mathbb{R}$, which by Bochner's theorem corresponds to a finite nonnegative Borel measure Λ. In that case, we have

$$\|\mu(p) - \mu(q)\| = \|F^{-1}[(\bar{\phi_p} - \bar{\phi_q})\Lambda]\|,$$

where ϕ_p is the characteristic function of the measure p, $\|.\|$ is the norm of the RKHS, F^{-1} is the inverse Fourier transform, and the bar denotes complex conjugation. Roughly speaking, this shows that p and q can be distinguished as long as the spectrum Λ of the kernel is nonzero wherever the spectra of the distributions might differ. If $\mathrm{supp}(\Lambda) = \mathbb{R}^d$, the kernel can distinguish all Borel distributions; if $\mathrm{supp}(\Lambda) \subset \mathbb{R}^d$ has a non-empty interior, it can still distinguish Borel distributions with compact support, subject to certain technical conditions (for details, see [5]).

The map μ has applications in a number of tasks including testing of homogeneity and independence [2, 3]. One can also establish a link to wave optics, which we will briefly sketch presently. We consider p as the intensity distribution of the light coming from an object which we would like to image. On the way to the sensor, there is an aperture with indicator function L (i.e., L takes the value 1 in the aperture, and 0 elsewhere). In the setting of Fraunhofer diffraction, the image intensity arising from a point source is the squared Fourier transform of L, i.e., the Fourier transform of the convolution of L with itself, $\Lambda := L * L$. For instance, in the 1-D case, if L is the indicator function of an interval, then Λ is a B_1-spline. Under the assumption of incoherent light, the image of p would thus be the convolution of p with the Fourier transform of Λ, equalling the map $\mu(p)$ induced by the translation invariant kernel associated with the Fourier transform of Λ. If the image has compact support, and the aperture has non-empty interior, then the imaging process is thus invertible.

References

[1] K. Fukumizu, A. Gretton, X. Sun and B. Schölkopf, *Kernel measures of conditional dependence*, Advances in Neural Information Processing Systems 20, pp. 489-496, MIT Press, Cambridge, MA, USA (2008).

[2] A. Gretton, K. M. Borgwardt, M. Rasch, B. Schölkopf and A. Smola, *A kernel method for the Two-Sample-Problem*, Advances in Neural Information Processing Systems 19, (2007).

[3] A. Gretton, K. Fukumizu, C. H. Teo, L. Song, B. Schölkopf and A. Smola, *A kernel statistical test of independence*, Advances in Neural Information Processing Systems 20, (2008).

[4] B. Schölkopf and A. J. Smola, *Learning with Kernels*, MIT Press, Cambridge, MA, USA (2002).

[5] B. K. Sriperumbudur, A. Gretton, K. Fukumizu, G. Lanckriet and B. Schölkopf, *Injective Hilbert space embeddings of probability measures*, Proceedings of the 21st Annual Conference on Learning Theory, USA (2008).

Multivariate Bernstein Basis Polynomials and their Kernels I

Kurt Jetter

(joint work with Elena E. Berdysheva, Joachim Stöckler)

The results addressed in this talk partially support the idea to extend the estimates given in [5] or in [6] to cases where multivariate approximation schemes are used and where polynomial kernels for SVM classification algorithms are built from good conditioned polynomial bases. The Bernstein basis polynomials are known to be much better conditioned than the monomial basis [4], and it is expected

that they will be useful in order to make some bias-variance estimates in learning theory more efficient and more transparent.

Let

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

denote the standard simplex in $\mathbb{R}^d$. The Bernstein basis polynomials of degree n in d variables, using barycentric coordinates

$$\mathbf{x} = (x_0, x_1, \ldots, x_d) \ , \quad x_0 = 1 - x_1 - \cdots - x_d \ ,$$

are defined by

$$B_\alpha(\mathbf{x}) = \frac{n!}{\alpha_0! \alpha_1! \ldots \alpha_d!} \, x_0^{\alpha_0} x_1^{\alpha_1} \cdots x_d^{\alpha_d} \ , \quad |\alpha| = n \ ,$$

for $\alpha = (\alpha_0, \alpha_1, \ldots, \alpha_d) \in \mathbb{N}_0^{d+1}$ and $|\alpha| = \alpha_0 + \alpha_1 + \cdots + \alpha_d$. They form a basis of Π_n^d, the space of d-variate polynomials of total degree at most n.

We consider here kernels of the form

$$\mathbf{T}_{n,\omega}(\mathbf{x}, \mathbf{y}) = \sum_{|\alpha|=n} \omega_\alpha \, B_\alpha(\mathbf{x}) \, B_\alpha(\mathbf{y}) \ , \quad \mathbf{x}, \mathbf{y} \in \mathbb{S}^d \ , \tag{1}$$

for positive weights $\omega = \big(\omega_\alpha\big)_{|\alpha|=n}$, and the corresponding integral operators are given by

$$(\mathbf{L}_{\rho,n,\omega} f)(\mathbf{x}) = \int_{\mathbb{S}^d} \mathbf{T}_{n,\omega}(\mathbf{x}, \mathbf{y}) \, f(\mathbf{y}) \, d\rho(\mathbf{y}) \ , \tag{2}$$

for f from an appropriate space of functions defined on the simplex. Here, ρ is assumed to be a non-negative measure, whence the operator is positive. For special situations (see [1]), kernels of this type have the remarkable analytical property that the sequence $\big(\mathbf{T}_{n,\omega}(\mathbf{x}, \mathbf{y})\big)_{n\in\mathbb{N}}$, for fixed $\mathbf{x}, \mathbf{y} \in \mathbb{S}^d$, is completely monotone. In case the weights are chosen so as to enforce the integral operator to reproduce constant functions, $\mathbf{L}_{\rho,n,\omega}$ is the so-called Bernstein-Durrmeyer operator, see [2, 3].

Since the polynomial kernel $\mathbf{T}_{n,\omega}(\mathbf{x}, \mathbf{y})$ is a Mercer kernel, we may look at the polynomial space Π_n^d as a RKHS (reproducing kernel Hilbert space) $H_{n,\omega}$ where the norm is defined via the semi-innerproduct

$$\left\langle \sum_{i=1}^N c_i \mathbf{K}_{\mathbf{x}_i} \mid \sum_{j=1}^N d_j \mathbf{K}_{\mathbf{x}_j} \right\rangle_{H_{n,\omega}} = \sum_{i,j=1}^N c_i \, \mathbf{T}_{n,\omega}(\mathbf{x}_i, \mathbf{x}_j) \, d_j \ ,$$

for given points $\mathbf{x}_i \in \mathbb{S}^d$. Here, $\mathbf{x} \mapsto \mathbf{K}_\mathbf{x} = \mathbf{T}_{n,\omega}(\mathbf{x}, \cdot)$ is the so-called feature map. With respect to this semi-innerproduct, the Bernstein basis polynomials are orthogonal,

$$\left\langle \sqrt{\omega_\alpha} B_\alpha \mid \sqrt{\omega_\beta} B_\beta \right\rangle_{H_{n,\omega}} = \delta_{\alpha,\beta} \ , \quad \text{for } |\alpha| = |\beta| = n \ .$$

This latter property can be used in order to estimate the norm of the operator $\mathbf{L}_{\rho,n,\omega}$, considered as an operator from the space $C(\mathbb{S}^d)$ into the space $H_{n,\omega}$. We

have - with **1** the constant function taking the value 1 -

$$\|\mathbf{L}_{\rho,n,\omega}\|^2_{C(\mathbb{S}^d)\to H_{n,\omega}} = \|\mathbf{L}_{\rho,n,\omega}\mathbf{1}\|^2_{H_{n,\omega}} = \sum_{|\alpha|=n} \omega_\alpha \left(\int_{\mathbb{S}^d} B_\alpha(\mathbf{y})\, d\rho(\mathbf{y})\right)^2 ,$$

and the right-hand side can be bounded from below and from above by

$$c_{min}\, M_\rho \le \|\mathbf{L}_{\rho,n,\omega}\|^2_{C(\mathbb{S}^d)\to H_{n,\omega}} \le c_{max}\, M_\rho , \tag{3}$$

with $M_\rho = \int_{\mathbb{S}^d} d\rho(\mathbf{y})$ the total mass of the measure, and c_{min}, c_{max} the minimal and maximal entry from the coefficients

$$c_{n,\rho}(\alpha) = \omega_\alpha \int_{\mathbb{S}^d} B_\alpha(\mathbf{y})\, d\rho(\mathbf{y}) , \quad |\alpha| = n .$$

For the Bernstein-Durrmeyer operator, these coefficients are all equal 1, but for this case the weights ω_α are intimately connected with the measure ρ. Estimate (3) allows us to get away from this restriction while keeping a control on the norm of the operator.

References

[1] E. Berdysheva, *Multivariate Bernstein Basis Polynomials and their Kernels II. Jacobi Weights*, this report.

[2] E. Berdysheva, K. Jetter and J. Stöckler, *New polynomial preserving operators on simplices: direct results*, J. Approximation Theory **131** (2004), 59–73.

[3] E. Berdysheva, K. Jetter and J. Stöckler, *Durrmeyer operators and their natural quasi-interpolants*, in: Topics in Multivariate Approximation and Interpolation (K. Jetter et al., eds.), pp. 1–21, Elsevier, Amsterdam, 2006.

[4] T. Lyche and K. Scherer, *On the p-norm condition number of the multivariate triangular Bernstein basis*, J. Comp. Appl. Math. **119** (2000), 259–273.

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

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

Multivariate Bernstein Basis Polynomials and their Kernels II: Jacobi Weights

Elena E. Berdysheva

(joint work with Kurt Jetter, Joachim Stöckler)

In this talk we give a survey on the Bernstein-Durrmeyer operators with Jacobi weights. These operators are special cases of a more general construction discussed in K. Jetter's talk [7].

Let

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

denote the standard simplex in $\mathbb{R}^d$. The *Bernstein basis polynomials* of degree n are defined by the formula

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

where $\alpha = (\alpha_0, \alpha_1, \dots, \alpha_d) \in \mathbb{N}_0^{d+1}$ with $|\alpha| = \alpha_0 + \cdots + \alpha_d = n$. The Jacobi weight is defined by $\omega_\mu(x) = (1 - x_1 - \cdots - x_d)^{\mu_0} x_1^{\mu_1} \cdots x_d^{\mu_d}$ for $\mu = (\mu_0, \mu_1, \dots, \mu_d) \in \mathbb{R}^{d+1}$ with $\mu_i > -1$, $i = 0, \dots, d$. We put $|\mu| = \mu_0 + \mu_1 + \cdots + \mu_d$ and $\underline{\mu} = \min\limits_{0 \le i \le d} \mu_i$. In a standard way, we define the weighted inner product $\langle f, g \rangle_\mu$ and the spaces $L^p_\mu(\mathbb{S}^d)$. Let $\mathbf{P}_n$ denote the space of algebraic polynomials of total degree at most n. Let $\mathbf{E}_{0,\mu} = \mathbf{P}_0$ and $\mathbf{E}_{m,\mu} = \{p \in \mathbf{P}_m : \langle p, g \rangle_\mu = 0 \ \forall g \in \mathbf{P}_{m-1}\}$, $m \in \mathbb{N}$, be the corresponding spaces of orthogonal polynomials.

The *Bernstein-Durrmeyer operator* is defined by

$$\mathbf{M}_{n,\mu} f = \sum_{|\alpha|=n} \frac{\langle f, B_\alpha \rangle_\mu}{\langle \mathbf{1}, B_\alpha \rangle_\mu} B_\alpha \tag{1}$$

for $f \in L^p_\mu(\mathbb{S}^d)$, $1 \le p < \infty$, or $f \in C(\mathbb{S}^d)$. Here, $\mathbf{1}$ denotes the constant function equal to one. Operator (1) was introduced in the one-dimensional unweighted case by Durrmeyer in 1967 and, independently, by Lupaş in 1972, and became popular after works by Derriennic [4]. The Bernstein-Durrmeyer operator with Jacobi weights was introduced by Păltănea and studied by Berens and Xu [3]; the multidimensional theory is due to Derriennic [5], and Ditzian [6]. Operator (1) was studied by many authors.

The Bernstein-Durrmeyer operator has the following properties:

(1) it is positive, i.e. $\mathbf{M}_{n,\mu} f \ge 0$ if $f \ge 0$,
(2) $\|\mathbf{M}_{n,\mu} f\|_{p,\mu} \le \|f\|_{p,\mu}$, $1 \le p \le \infty$,
(3) it reproduces constant functions, i.e. $\mathbf{M}_{n,\mu} p = p$ for $p \in \mathbf{P}_0$,
(4) it is degree reducing, i.e. $p - \mathbf{M}_{n,\mu} p \in \mathbf{P}_{m-1}$ for $p \in \mathbf{P}_m$, $m \le n$,
(5) it is self-adjoint, i.e. $\langle \mathbf{M}_{n,\mu} f, g \rangle_\mu = \langle f, \mathbf{M}_{n,\mu} g \rangle_\mu$.

Theorem 1. [4, 5, 3, 6] *For all $n \in \mathbb{N}$, the spaces $\mathbf{E}_{m,\mu}$, $m \in \mathbb{N}_0$, are eigenspaces of the Bernstein-Durrmeyer operator, and*

$$\mathbf{M}_{n,\mu} p_m = \gamma_{n,m,\mu} \ p_m \quad \textit{for} \quad p_m \in \mathbf{E}_{m,\mu},$$

where $\gamma_{n,m,\mu} = \binom{n}{m} / \binom{n+d+|\mu|+m}{m}$.

This nice spectral decomposition as well as the property of degree reduction are not valid in the general case considered in [7].

The operator $\mathbf{M}_{n,\mu}$ is an integral operator

$$(\mathbf{M}_{n,\mu} f)(y) = \int_{\mathbb{S}^d} K_{n,\mu}(x, y) \, f(x) \, \omega_\mu(x) \, dx,$$

with the kernel function

$$K_{n,\mu}(x,y) = \sum_{|\alpha|=n} \frac{1}{\langle \mathbf{1}, B_\alpha\rangle_\mu} B_\alpha(x)\, B_\alpha(y).$$

We consider the rescaled kernel

$$T_{n,\mu}(x,y) = \frac{\Gamma(n+\underline{\mu}+1)}{\Gamma(n+|\mu|+d+1)} K_{n,\mu}(x,y).$$

Theorem 2. [2] *Let $\mu \in \mathbb{R}^{d+1}$ be such that $\underline{\mu} \ge -\frac{1}{2}$. Then, for every $x, y \in \mathbb{S}^d$, the sequence $(T_{n,\mu}(x,y))_{n\in\mathbb{N}}$ is bounded and completely monotonic; i.e.*

$$(-1)^r \Delta^r T_{n,\mu}(x,y) = \sum_{\ell=0}^{r} (-1)^\ell \binom{r}{\ell} T_{n+\ell,\mu}(x,y) \ge 0, \qquad r, n \ge 0.$$

The differential operator

$$\begin{aligned} -\mathbf{U}_\mu \;&=\; \sum_{i=1}^{d} (\omega_\mu(x))^{-1} \frac{\partial}{\partial x_i} \left\{ \omega_\mu(x)(1 - x_1 - \cdots - x_d) x_i \frac{\partial}{\partial x_i} \right\} \\ &+\; \sum_{1\le i<j\le d} (\omega_\mu(x))^{-1} \left(\frac{\partial}{\partial x_j} - \frac{\partial}{\partial x_i} \right) \left\{ \omega_\mu(x) x_j x_i \left(\frac{\partial}{\partial x_j} - \frac{\partial}{\partial x_i} \right) \right\} \end{aligned}$$

plays an important role in the analysis of the operator $\mathbf{M}_{n,\mu}$. In particular [4, 5, 3, 6], the following Voronovskaya type statement holds true: if $f \in C^2(\mathbb{S}^d)$, then

$$\lim_{n\to\infty} n\{f - \mathbf{M}_{n,\mu} f\}(x) = \mathbf{U}_\mu f(x). \tag{2}$$

Being a positive operator, $\mathbf{M}_{n,\mu}$ cannot converge fast, compare (2). To accelerate the convergence, Jetter and Stöckler [8] introduced the *natural quasi-interpolants of Bernstein-Durrmeyer operators* of order (r,n), $0 \le r \le n$,

$$\mathbf{M}_{n,\mu}^{(r)} f = \sum_{\ell=0}^{r} \frac{1}{\binom{n}{\ell}} \mathbf{U}_{\ell,\mu}(\mathbf{M}_{n,\mu} f).$$

The differential operators $\mathbf{U}_{\ell,\mu}$ here are defined recursively: $\mathbf{U}_{0,\mu} = \mathbf{I}$, the identity operator, and

$$\mathbf{U}_{\ell+1,\mu} = \frac{1}{(\ell+1)^2} (\mathbf{U}_\mu - \ell(\ell+d+|\mu|)\mathbf{I}) \mathbf{U}_{\ell,\mu}, \quad \ell \in \mathbb{N}.$$

The quasi-interpolants $\mathbf{M}_{n,\mu}^{(r)}$ have the following properties [8, 1, 2]:

(1) $\mathbf{M}_{n,\mu}^{(0)} = \mathbf{M}_{n,\mu}$ and $\mathbf{M}_{n,\mu}^{(n)}\Big|_{\mathbf{P}_n} = \mathbf{I}|_{\mathbf{P}_n}$,

(2) they reproduce polynomials: $\mathbf{M}_{n,\mu}^{(r)}(p) = p$ for $p \in \mathbf{P}_r$, $0 \le r \le n$,

(3) $\mathbf{M}_{n,\mu}^{(r)}$ are bounded uniformly in n.

It was shown in [8] that the spaces $\mathbf{E}_{m,\mu}$, $m \in \mathbb{N}_0$, are eigenspaces of the operators $\mathbf{U}_{\ell,\mu}$, $\ell \in \mathbb{N}$. It follows that the spaces $\mathbf{E}_{m,\mu}$ are also eigenspaces of the operator $\mathbf{M}_{n,\mu}^{(r)}$. The approximation behavior of $\mathbf{M}_{n,\mu}^{(r)}$ is described in the following

theorems. These statements were obtained using the spectral properties of the involved operators and the complete monotonicity property (Theorem 2).

Theorem 3. [2] *The operators* $\mathbf{M}_{n,\mu}^{(r)}$ *can be represented as linear combinations of the Durrmeyer operators:*

$$\mathbf{M}_{n,\mu}^{(r)} = \sum_{\ell=0}^{r} (-1)^{\ell} \binom{r}{\ell} \binom{n+d+|\mu|+r-\ell}{r} \mathbf{M}_{n-\ell,\mu}.$$

Theorem 4. [1] *For* $f \in C^{2r+2}(\mathbb{S}^d)$*, the following Voronovskaya type result holds true:*

$$\lim_{n\to\infty} \binom{n}{r+1} \{f(x) - (\mathbf{M}_{n,\mu}^{(r)} f)(x)\} = (\mathbf{U}_{r+1,\mu} f)(x).$$

Theorem 5. [1] *For* $f \in C^{2r+2}(\mathbb{S}^d)$*, the following Jackson-Favard type estimate holds true:*

$$\|f - \mathbf{M}_{n,\mu}^{(r)} f\|_{p,\mu} \le \frac{C_{r,d,\mu}}{\binom{n}{r+1}} \|\mathbf{U}_{r+1,\mu} f\|_{p,\mu}.$$

Thus, the rate of convergence for smooth functions is n^{-r-1}. To deal with non-smooth functions, we introduce for $f \in L_\mu^p(\mathbb{S}^d)$, $1 \le p \le \infty$, a new K-functional

$$K_{\ell,p,\mu}(f,t) := \inf_{g \in C^{2\ell}(\mathbb{S}^d)} \{\|f-g\|_{p,\mu} + t\,\|\mathbf{U}_{\ell,\mu}(g)\|_{p,\mu}\}.$$

Theorem 6. [2] *Let* $n, r \in \mathbb{N}_0$*,* $0 \le r \le n$*, and* $\mu \in \mathbb{R}^{d+1}$ *with* $\underline{\mu} \ge -\frac{1}{2}$*. Then*

$$\|f - \mathbf{M}_{n,\mu}^{(r)} f\|_{p,\mu} \le C_{r,\mu,d}\, K_{r+1,p,\mu}(f, n^{-r-1}).$$

References

[1] E. Berdysheva, K. Jetter and J. Stöckler, *New polynomial preserving operators on simplices: direct results*, J. Approximation Theory **131** (2004), 59–73.

[2] E. Berdysheva, K. Jetter and J. Stöckler, *Durrmeyer operators and their natural quasi-interpolants*, in: Topics in Multivariate Approximation and Interpolation (K. Jetter et al., eds.), pp. 1–21, Elsevier, Amsterdam, 2006.

[3] H. Berens and Y. Xu, *On Bernstein-Durrmeyer polynomials with Jacobi-weights*, in: Approximation Theory and Functional Analysis (C. K. Chui, Ed.), pp. 25–46, Academic Press, Boston, 1991.

[4] M.-M. Derriennic, *Sur l'approximation de fonctions intégrables sur* $[0,1]$ *par des polynômes de Bernstein modifies*, J. Approximation Theory **31** (1981), 325–343.

[5] M.-M. Derriennic, *On multivariate approximation by Bernstein-type polynomials*, J. Approximation Theory **45** (1985), 155–166.

[6] Z. Ditzian, *Multidimensional Jacobi-type Bernstein-Durrmeyer operators*, Acta Sci. Math. (Szeged) **60** (1995), 225–243.

[7] K. Jetter, *Multivariate Bernstein Basis Polynomials and their Kernels I*, this report.

[8] K. Jetter and J. Stöckler, *An identity for multivariate Bernstein polynomials*, Computer Aided Geometric Design **20** (2003), 563–577.

Reporter: Kurt Jetter

Participants

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

Prof. Dr. Peter G. Binev
Dept. of Mathematics
University of South Carolina
Columbia, SC 29208
USA

Prof. Dr. Stephane Boucheron
UFR de Mathematiques
Universite Denis Diderot-Paris 7
175, rue du Chevaleret
F-75013 Paris

Prof. Dr. Andrea Caponnetto
DISI
Universita di Genova
V. Dodecaneso 35
I-16146 Genova

Prof. Dr. Bin Han
Department of Mathematical and
Statistical Sciences
University of Alberta
541 Central Academic Building
Edmonton 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
70593 Stuttgart

Prof. Dr. Michael Kohler
Fachbereich Mathematik
TU Darmstadt
Schloßgartenstr. 7
64289 Darmstadt

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

Dr. Massimiliano Pontil
Department of Computer Science
University College London
Gower Street
GB-London WC1E 6BT

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

Prof. Dr. Tomas Sauer
Lehrstuhl für Numerische Mathematik
Universität Gießen
Heinrich-Buff-Ring 44
35392 Giessen

Prof. Dr. Bernhard Schölkopf
Max-Planck-Institut für
Biologische Kybernetik
Spemannstraße 38
72076 Tübingen

Prof. Dr. Steve Smale
Department of Mathematics
University of California
Berkeley, CA 94720-3840
USA

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

Prof. Dr. Johan Suykens
Dept. of Electrical Engineering
K.U. Leuven
Kasteelpark Arenberg 10
B-3001 Leuven

Prof. Dr. Pierre Tarres
St. Hugh's College
St. Margaret's Road
GB-Oxford OX2 6LE

Prof. Dr. Vladimir N. Temlyakov
Dept. of Mathematics
University of South Carolina
Columbia, SC 29208
USA

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

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

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