Nonparametric Regression Based on Hierarchical Interaction Models

Michael Kohler and Adam Krzyżak, *Fellow, IEEE*

***Abstract*— In this paper, we introduce the so-called hierarchical interaction models, where we assume that the computation of the value of a function $m : \mathbb{R}^d \to \mathbb{R}$ is done in several layers, where in each layer a function of at most d^* inputs computed by the previous layer is evaluated. We investigate two different regression estimates based on polynomial splines and on neural networks, and show that if the regression function satisfies a hierarchical interaction model and all occurring functions in the model are smooth, the rate of convergence of these estimates depends on d^* (and not on d). Hence, in this case, the estimates can achieve good rate of convergence even for large d, and are in this sense able to circumvent the so-called curse of dimensionality.**

***Index Terms*— Curse of dimensionality, dimension reduction, interaction models, L_2 error, nonparametric regression, projection pursuit, rate of convergence.**

I. Introduction

In regression analysis a random vector (X, Y) with values in $\mathbb{R}^d \times \mathbb{R}$ satisfying $\mathbf{E}Y^2 < \infty$ is given and the goal is to predict the value of response variable Y given the value of observation vector X. If the main aim of the analysis is minimization of the mean squared error or L_2 risk, then a function $m^* : \mathbb{R}^d \to \mathbb{R}$ is sought satisfying

$$\mathbf{E}\{|Y - m^*(X)|^2\} = \min_{f:\mathbb{R}^d \to \mathbb{R}} \mathbf{E}\{|Y - f(X)|^2\}.$$

Let $m : \mathbb{R}^d \to \mathbb{R}$, $m(x) = \mathbf{E}\{Y|X = x\}$ be the regression function. Since

$$\mathbf{E}\{|Y - f(X)|^2\} = \mathbf{E}\{|Y - m(X)|^2\} + \int |f(x) - m(x)|^2 \mathbf{P}_X(dx)$$

(cf., e.g., Györfi *et al.* [9, Sec. 1.1]), the regression function is the optimal predictor $m^* = m$, and any function $f : \mathbb{R}^d \to \mathbb{R}$ is a good predictor in the sense that its L_2 risk is close to the optimal value if and only if the L_2 error

$$\int |f(x) - m(x)|^2 \mathbf{P}_X(dx)$$

is small. In nonparametric regression a set of data

$$\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$$

is given, where (X, Y), (X_1, Y_1), (X_2, Y_2), …are independent and identically distributed random variables, and the aim is to construct a regression estimate $m_n(\cdot) = m_n(\cdot, \mathcal{D}_n)$ such that its L_2 error

$$\int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx)$$

is small.

To appreciate the difference between parametric and nonparametric regression estimation note that in parametric estimation one assumes that the structure of the regression function is known and depends only on finitely many parameters, and one uses the data to estimate the (unknown) values of these parameters. In nonparametric approach we do not assume that the regression function can be described by finitely many parameters and the whole function is estimated from the data.

For a systematic and rigorous coverage of nonparametric regression estimation refer to Györfi *et al.* [9].

In order to derive nontrivial results on the rate of convergence of the expected L_2 error, it is necessary to impose smoothness assumptions on m (cf., e.g., Györfi *et al.* [9, Th. 3.1]). It was shown in Stone [25] that the optimal minimax rate of convergence for estimation of (p, C)-smooth regression function (where roughly speaking, see below for the exact definition, the regression function is p-times continuously differentiable) is

$$n^{-\frac{2p}{2p+d}}$$

where the minimax rate of convergence is defined as follows:

The sequence of (eventually) positive numbers a_n is called a **lower minimax rate of convergence** for the class $\mathcal{D}$ if

$$\liminf_{n\to\infty} \inf_{m_n} \sup_{(X,Y)\in\mathcal{D}} \frac{\mathbf{E}\{\|m_n - m\|^2\}}{a_n} = C_1 > 0.$$

The sequence is said to be an **achievable rate of convergence** for the class $\mathcal{D}$ if

$$\limsup_{n\to\infty} \sup_{(X,Y)\in\mathcal{D}} \frac{\mathbf{E}\{\|m_n - m\|^2\}}{a_n} = C_2 < \infty.$$

Manuscript received June 11, 2015; revised May 30, 2016; accepted November 7, 2016. Date of publication December 1, 2016; date of current version February 14, 2017. This work was supported in part by the German Research Foundation within the Collaborative Research Centre 805 and in part by the Natural Sciences and Engineering Research Council of Canada under Grant RGPIN-2015-06412 and the authors would like to thank DFG and NSERC for funding this work. This paper was presented at the 2016 IEEE International Symposium on Information Theory. *(Corresponding author: Adam Krzyżak.)*

M. Kohler is with the Fachbereich Mathematik, Technische Universität Darmstadt, 64289 Darmstadt, Germany (e-mail: kohler@mathematik.tu-darmstadt.de).

A. Krzyżak is with the Department of Computer Science and Software Engineering, Concordia University, Montreal, QC H3G 1M8, Canada (e-mail: krzyzak@cs.concordia.ca).

Communicated by T. Javidi, Associate Editor for Communication Networks.
Digital Object Identifier 10.1109/TIT.2016.2634401

The sequence is called an *optimal minimax rate of convergence* if it is both a lower minimax and an achievable rate of convergence.

If d is large compared to p, then this rate of convergence is rather slow, which is a consequence of the fact, that high-dimensional regression problems are especially difficult to solve due to so–called curse of dimensionality. But unfortunately, most applications are high-dimensional problems and hence very hard to solve. The only way to circumvent this curse of dimensionality is to impose additional assumptions on the regression function in order to derive better rates of convergence.

Stone [26] proposed to impose an additivity condition on the structure of the regression function. He assumed that

$$m(x^{(1)}, \dots, x^{(d)}) = m_1(x^{(1)}) + \cdots + m_d(x^{(d)}) \quad (x = (x^{(1)}, \dots, x^{(d)})^T \in \mathbb{R}^d)$$

for (p, C)-smooth univariate functions $m_1, \dots, m_d : \mathbb{R} \to \mathbb{R}$, and he was able to show that in this case $n^{-2p/(2p+1)}$ is the optimal minimax rate of convergence. A generalization of this approach to so–called interaction models was presented in Stone [27]. Here it was assumed that for some $d^* \in \{1, \dots, d\}$ the regression function satisfies

$$m(x) = \sum_{I \subseteq \{1,\dots,d\}, |I| = d^*} m_I(x_I),$$

where $|I|$ denotes the cardinality of the set I, m_I are (p, C)-smooth functions defined on $\mathbb{R}^{|I|}$ and for $x = (x^{(1)}, \dots, x^{(d)})^T$ and $I = \{i_1, \dots, i_{d^*}\}$ with $1 \le i_1 < \cdots < i_{d^*} \le d$ we set $x_I = (x^{(i_1)}, \dots, x^{(i_{d^*})})^T$. In other words, it is assumed that the regression function is a sum of (p, C)-smooth functions where each function in the sum depends on at most d^* of the components of x. Under this assumption it was shown that $n^{-2p/(2p+d^*)}$ is the optimal minimax rate of convergence.

Other models which yield good rates of convergence even for high dimensional data include single index models and projection pursuit. In single index models it is assumed that

$$m(x) = g(a^T x) \quad (x \in \mathbb{R}^d),$$

where $g : \mathbb{R} \to \mathbb{R}$ is an univariate function and $a \in \mathbb{R}^d$ is a d-dimensional vector (cf., e.g., Härdle *et al.* [10], Härdle and Stoker [12], Kong and Xia [19], and Yu and Ruppert [29]). In projection pursuit the regression function is allowed to be a sum of functions of the above form, i.e.,

$$m(x) = \sum_{k=1}^{K} g_k(a_k^T x) \quad (x \in \mathbb{R}^d),$$

where $K \in \mathbb{N}$ is a natural number, $g_k : \mathbb{R} \to \mathbb{R}$ are univariate functions and $a_k \in \mathbb{R}^d$ are d-dimensional vectors (cf., e.g., Friedman and Stuetzle [8]). If the univariate functions above are (p, C)-smooth, then corresponding regression estimates can achieve under these assumptions the corresponding univariate rates of convergence (cf., e.g., Györfi et al. [9, Ch. 22]).

Horowitz and Mammen [16] considered a general regression model

$$m(x) = g\left(\sum_{l_1=1}^{L_1} g_{l_1}\left(\sum_{l_2=1}^{L_2} g_{l_1,l_2}\left(\cdots \sum_{l_p=1}^{L_p} g_{l_1,\dots,l_p}(x^{l_1,\dots,l_p})\right)\right)\right)$$

where $g, g_{l_1}, \dots, g_{l_1,\dots,l_p}$ are assumed to be unknown (p, C)-smooth univariate functions and $x^{l_1,\dots,l_p}$ are one-dimensional elements of a covariate vector x, which may be identical for two different indices $(l_1, \dots, l_p)$. They estimated the model by the penalized least squares and obtained the rate $n^{-2p/(2p+1)}$. Their model is more restrictive than our generalized hierarchical interaction model introduced in Section 2. In addition they work with smoothing splines whereas we estimate our model with multilayer neural networks.

A mixture of parametric and nonparametric approach is achieved in semiparametric models. Here it is assumed that for a part of the components of x the influence on the regression function is known and is described by a parametric model (e.g., a linear model), and only the remaining part is estimated nonparametrically (cf., e.g., Härdle *et al.* [11]). Under this assumption the corresponding estimates are able to achieve rates of convergence corresponding to d^* dimensional problems, where d^* is the number of components of x for which a parametric model is not given.

In any application these estimates achieve good rates of convergence only if the imposed assumptions are satisfied. Our research in this paper is motivated by applications in connection with complex technical systems, which are constructed in a modular form (in particular a load bearing structure studied currently by the Collaborative Research Centre 805 at the Technische Universität Darmstadt, Germany). If such systems are constructed in a modular form, then it is plausible to model the outcome of the system as a function of the outputs of the modular parts of it, where each modular part computes a function depending only on a subset of the components of the high-dimensional input. In the simplest case we formulate this by assuming that our regression function satisfies

$$m(x) = g(f_1(x_{I_1}), \dots, f_{d^*}(x_{I_{d^*}})) \quad (x \in \mathbb{R}^d),$$

where $d^* \in \{1, \dots d\}$, $I_1, \dots, I_{d^*} \subseteq \{1, \dots, d\}$ are sets of cardinality d^* and $g, f_1, \dots, f_{d^*}$ are (p, C)-smooth functions defined on $\mathbb{R}^{d^*}$. The corresponding general case can be found in Section 2. In the general case the above model is recursively applied, which is reasonable especially if we consider a complex technical system constructed in a modular form, where each modular part may be again a complex system constructed in a similar modular form. Under the above assumption we show that suitably defined spline and neural network estimates achieve (up to some logarithmic factor) the rate of convergence $n^{-2p/(2p+d^*)}$.

A. Notation

Throughout the paper the following notation is used: The sets of positive integers, nonnegative integers, integers, nonnegative real numbers and real numbers are denoted by $\mathbb{N}$,

$\mathbb{N}_0$, $\mathbb{Z}$, $\mathbb{R}_+$ and $\mathbb{R}$, resp. Let $D \subseteq \mathbb{R}^d$ and let $f : \mathbb{R}^d \to \mathbb{R}$ be a real-valued function defined on $\mathbb{R}^d$. We write $x = \arg\max_{z \in D} f(z)$ if $\max_{z \in D} f(z)$ exists and if x satisfies

$$x \in D \quad \text{and} \quad f(x) = \max_{z \in D} f(z).$$

The Euclidean and the supremum norms of $x \in \mathbb{R}^d$ are denoted by $\|x\|$ and $\|x\|_\infty$, resp. For $f : \mathbb{R}^d \to \mathbb{R}$

$$\|f\|_\infty = \sup_{x \in \mathbb{R}^d} |f(x)|$$

is its supremum norm, and the supremum norm of f on a set $A \subseteq \mathbb{R}^d$ is denoted by

$$\|f\|_{\infty, A} = \sup_{x \in A} |f(x)|.$$

The support of an $\mathbb{R}^d$–valued random variable X is abbreviated by

$$supp(X) = \left\{ x \in \mathbb{R}^d \, : \, \mathbf{P}_X(S_r(x)) > 0 \text{ for all } r > 0 \right\},$$

where $S_r(x)$ is the ball of radius r around x.

The outline of this paper is as follows: The assumption on the structure of the regression function is described in Section 2, in Section 3 we introduce estimates based on polynomial splines and present a result concerning their rates of convergence, Section 4 does the same for neural networks, and Section 5 contains the proofs.

II. Hierarchical Interaction Models

In this section we formalize the assumption that a function value is computed in several layers where in each layer a function of at most d^* inputs is computed and where the inputs are outputs of the previous layer (or components of the input variable, in case that there is no previous layer). We do this in the following recursive definition.

Definition 1: Let $d \in \mathbb{N}$, $d^* \in \{1, \dots, d\}$, $l \in \mathbb{N}_0$ and $m : \mathbb{R}^d \to \mathbb{R}$.

a) *We say that m satisfies a* **hierarchical interaction model of order** d^* **and level** 0*, if there exist $I \subseteq \{1, \dots, d\}$ with $|I| = d^*$ and $f : \mathbb{R}^{d^*} \to \mathbb{R}$ such that*

$$m(x) = f(x_I) \quad \text{for all } x \in \mathbb{R}^d.$$

b) *We say that m satisfies a* **hierarchical interaction model of order** d^* **and level** $l+1$*, if there exist $g : \mathbb{R}^{d^*} \to \mathbb{R}$ and $f_1, \dots, f_{d^*} : \mathbb{R}^d \to \mathbb{R}$ such that $f_1, \dots, f_{d^*}$ satisfy a hierarchical interaction model of order d^* and level l and such that*

$$m(x) = g\left(f_1(x), \dots, f_{d^*}(x)\right) \quad \text{for all } x \in \mathbb{R}^d.$$

The class of functions satisfying a hierarchical interaction model of order 1 neither includes all additive functions nor all functions satisfying the assumption of projection pursuit. But after a slight extension of the definition, which we present next, all such functions are included.

Definition 2: Let $d \in \mathbb{N}$, $d^* \in \{1, \dots, d\}$, $l \in \mathbb{N}_0$ and $m : \mathbb{R}^d \to \mathbb{R}$.

a) *We say that m satisfies a* **generalized hierarchical interaction model of order** d^* **and level** 0*, if there exist $a_1, \dots, a_{d^*} \in \mathbb{R}^{d^*}$ and $f : \mathbb{R}^{d^*} \to \mathbb{R}$ such that*

$$m(x) = f(a_1^T x, \dots, a_{d^*}^T x) \quad \text{for all } x \in \mathbb{R}^d.$$

b) *We say that m satisfies a* **generalized hierarchical interaction model of order** d^* **and level** $l+1$*, if there exist $K \in \mathbb{N}$, $g_k : \mathbb{R}^{d^*} \to \mathbb{R}$ $(k = 1, \dots, K)$ and $f_{1,k}, \dots, f_{d^*,k} : \mathbb{R}^d \to \mathbb{R}$ $(k = 1, \dots, K)$ such that $f_{1,k}, \dots, f_{d^*,k}$ $(k = 1, \dots, K)$ satisfy a generalized hierarchical interaction model of order d^* and level l and such that*

$$m(x) = \sum_{k=1}^{K} g_k\left(f_{1,k}(x), \dots, f_{d^*,k}(x)\right) \quad \text{for all } x \in \mathbb{R}^d.$$

Obviously, each hierarchical interaction model is also a generalized hierarchical interaction model of the same order and the same level (because we can choose $a_k's$ as unit vectors and $K = 1$). Furthermore, additive functions, all functions satisfying the assumption of projection pursuit or of interaction model belong to the class of generalized hierarchical interaction model of order d^* and level 1, where $d^* = 1$ in case of additive functions or projection pursuit.

Our smoothness assumptions imposed on the functions occurring in a hierarchical interaction model are formalized in the next definition.

Definition 3: **a)** *Let $p = k + s$ for some $k \in \mathbb{N}_0$ and $0 < s \leq 1$. A* **function** *$m : \mathbb{R}^d \to \mathbb{R}$ is called* **(p, C)-smooth***, if for every $\alpha = (\alpha_1, \dots, \alpha_d) \in \mathbb{N}_0^d$ with $\sum_{j=1}^d \alpha_j = k$ the partial derivative $\frac{\partial^k m}{\partial x_1^{\alpha_1} \dots \partial x_d^{\alpha_d}}$ exists and satisfies*

$$\left| \frac{\partial^k m}{\partial x_1^{\alpha_1} \dots \partial x_d^{\alpha_d}}(x) - \frac{\partial^k m}{\partial x_1^{\alpha_1} \dots \partial x_d^{\alpha_d}}(z) \right| \leq C \cdot \|x - z\|^s$$

for all $x, z \in \mathbb{R}^d$.

b) *We say that the* **(generalized) hierarchical interaction model** *in Definition 1 (Definition 2)* **is (p, C)–smooth***, if all functions occurring in its definition are (p, C)–smooth according to part a) of this definition.*

Remark 1: **a)** If

$$m(x) = g\left(f_1(x), \dots, f_{d^*}(x)\right) \quad (x \in \mathbb{R}^d)$$

for some (p, C)–smooth functions $g : \mathbb{R}^{d^*} \to \mathbb{R}$ and $f_1, \dots, f_{d^*} : \mathbb{R}^d \to \mathbb{R}$, then we get in case $p \leq 1$

$$\begin{aligned} &|m(x) - m(z)| \\ &\quad = |g\left(f_1(x), \dots, f_{d^*}(x)\right) - g\left(f_1(z), \dots, f_{d^*}(z)\right)| \\ &\quad \leq C \cdot \sqrt{d} \cdot \max_{j=1,\dots,d^*} |f_j(x) - f_j(z)|^p. \end{aligned}$$

Using this and

$$\|x_I - z_I\| \leq \|x - z\|$$

for $I \subseteq \{1, \dots, d\}$ and $x, z \in \mathbb{R}^d$ we see that for any $p \leq 1$ any function which satisfies a hierarchical interaction model of level l which is (p, C)–smooth according to Definition 3 b) is $(p^{l+1}, \bar{C})$-smooth according to Definition 3 a).

b) In the definition of (generalized) hierarchical interaction model it is possible to choose some of the occurring

functions as projections on some component of their input, which are always (p, C)–smooth functions. Consequently, (p, C)–smooth functions depending on at most d^* components of its input variable belong to the class of functions satisfying (p, C)–smooth (generalized) hierarchical interaction models of order d^* and any fixed level. Therefore we can conclude from Stone [25] that the minimax rate of convergence of estimation of (p, C)–smooth (generalized) hierarchical interaction models of order d^* is lower bounded by $n^{-2p/(2p+d^*)}$. In the next two sections we show that suitably defined spline and neural network estimates achieve this rate of convergence up to some logarithmic factor. In order to simplify the notation the result for splines is derived only for hierarchical interaction models, however the result for neural networks considers also generalized hierarchical interaction models.

III. Estimates Based on Polynomial Splines

In the next two sections we study least squares estimates defined by

$$m_n(\cdot) = \arg\min_{h \in \mathcal{H}_n} \frac{1}{n} \sum_{i=1}^{n} |Y_i - h(X_i)|^2, \tag{1}$$

where $\mathcal{H}_n$ is a set of functions $h : \mathbb{R}^d \to \mathbb{R}$. For simplicity we assume here and in the sequel that the minimum above indeed exists. When this is not the case our theoretical results also hold for any estimate which minimizes the above empirical L_2 risk up to a small additional term (e.g., $1/n$).

In this section we will define $\mathcal{H}_n$ by using tensor products of polynomial splines, i.e., tensor products of piecewise polynomials satisfying a global smoothness condition. Concerning applications of tensor products of polynomial splines in nonparametric regression we refer to Friedman [7], Kohler [17], and Stone *et al.* [28] and the literature cited therein.

In the sequel we introduce spaces of tensor product B–splines defined on $\mathbb{R}^d$ and then compose them according to the definition of hierarchical interaction models. Our function spaces will depend on parameters $\alpha > 0$ (controlling the supremum norm of the functions), $\beta > 0$ (controlling the support of the functions), $\gamma > 0$ (controlling the Lipschitz constant of the functions), $M_0 \in \mathbb{N}$ (the degree of the splines), $K \in \mathbb{N}$ (controlling the degrees of freedom) and d (the dimension of $\mathbb{R}^d$).

We start by introducing univariate space of polynomial spline functions and a corresponding B-spline basis consisting of basis functions with compact support as follows: For $K \in \mathbb{N}$ and $M \in \mathbb{N}_0$ set $u_k = k \cdot \beta/K$ $(k \in \mathbb{Z})$. For $k \in \mathbb{Z}$ let $B_{k,M} : \mathbb{R} \to \mathbb{R}$ be the univariate B-spline of degree M with knot sequence $(u_l)_{l \in \mathbb{Z}}$ and support $supp(B_{k,M}) = [u_k, u_{k+M+1}]$. In case $M = 0$ this means that $B_{k,0}$ is the indicator function of the interval $[u_k, u_{k+1})$, and for $M = 1$ we have

$$B_{k,1}(x) = \begin{cases} \dfrac{x - u_k}{u_{k+1} - u_k}, & u_k \le x \le u_{k+1}, \\ \dfrac{u_{k+2} - x}{u_{k+2} - u_{k+1}}, & u_{k+1} < x \le u_{k+2}, \\ 0, & \text{else}, \end{cases}$$

(so-called hat-function). The general definition of $B_{k,M}$ can be found, e.g., in de Boor [2], or in Györfi *et al.* [9, Sec. 14.1]. These B-splines are basis functions of sets of univariate piecewise polynomials of degree M, where the piecewise polynomials are globally $(M-1)$–times continuously differentiable and where the M-th derivative of the functions have jump points only at the knots u_l $(l \in \mathbb{Z})$.

For $\mathbf{k} = (k_1, \ldots, k_d) \in \mathbb{Z}^d$ we define the tensor product B-spline $B_{\mathbf{k},M} : \mathbb{R}^d \to \mathbb{R}$ by

$$B_{\mathbf{k},M}(x^{(1)}, \ldots, x^{(d)}) = B_{k_1,M}(x^{(1)}) \cdots B_{k_d,M}(x^{(d)}) \quad (x^{(1)}, \ldots, x^{(d)} \in \mathbb{R}).$$

And we define $\mathcal{S}_{K,M}$ as the set of all linear combinations of all those of the above tensor product B-splines, where the support has nonempty intersection with $(-\beta, \beta)^d$, i.e., we set

$$\mathcal{S}_{K,M} = \left\{ \sum_{\mathbf{k} \in \{-K-M, -K-M+1, \ldots, K-1\}^d} a_{\mathbf{k}} \cdot B_{\mathbf{k},M} \; : \; a_{\mathbf{k}} \in \mathbb{R} \right\}.$$

For our estimate we need to impose bounds on the supremum norm and the Lipschitz constant of our functions. We do this by restricting the coefficients in the spline space as follows: Let e_i be the i-th unit vector in $\mathbb{R}^d$ $(i = 1, \ldots, d)$. Then we set

$$\mathcal{S}_{K,M,\alpha,\beta,\gamma,d} = \left\{ \sum_{\mathbf{k} \in \mathbb{Z}^d} a_{\mathbf{k}} \cdot B_{\mathbf{k},M} \; : \; |a_{\mathbf{k}}| \le \alpha, |a_{\mathbf{k}} - a_{\mathbf{k}-e_i}| \le \frac{\beta \cdot \gamma}{\sqrt{d} \cdot K} \; (i = 1, \ldots, d), \; a_{\mathbf{k}} = 0 \text{ if } supp(B_{\mathbf{k},M}) \cap (-\beta, \beta)^d = \emptyset \right\}.$$

The definition of the B-splines implies that $\mathcal{S}_{K,M,\alpha,\beta,\gamma,d}$ is a subset of a linear vector space of dimension $(2 \cdot K + M)^d$. Furthermore, by standard results on B-splines and their derivatives (cf., e.g., Györfi *et al.* [9, Lemmas 14.4 and 14.6]) it can be shown that the functions in $\mathcal{S}_{K,M,\alpha,\beta,\gamma,d}$ are bounded in absolute value by α and are for $M > 0$ Lipschitz continuous with Lipschitz constant bounded by γ (since all partial derivatives of order one are bounded in absolute value by $\gamma/\sqrt{d}$).

Now we assume that we have given a hierarchical interaction model of order d^* and that we know all subsets I occurring in its definition. We use them to define similarly a composition of our spline spaces as follows:

For level 0 we define $\mathcal{H}^{(0)}$ by choosing $I \subseteq \{1, \ldots, d\}$ with $|I| = d^*$ and by setting

$$\mathcal{H}^{(0)} = \left\{ h : \mathbb{R}^d \to \mathbb{R} \; : \; h(x) = f(x_I) \; (x \in \mathbb{R}^d) \text{ for some } f \in \mathcal{S}_{K,M,\alpha,\beta,\gamma,d^*} \right\}.$$

For level $l+1$ we define $\mathcal{H}_1^{(l)}$, ..., $\mathcal{H}_{d^*}^{(l)}$ according to the functions chosen in the definition of our hierarchical interaction model of level l and set

$$\mathcal{H}^{(l+1)} = \left\{ h : \mathbb{R}^d \to \mathbb{R} \; : \; h(x) = g(f_1(x), \ldots, f_{d^*}(x)) \; (x \in \mathbb{R}^d) \text{ for some } g \in \mathcal{S}_{K,M,\alpha,\beta,\gamma,d^*}, \; f_1 \in \mathcal{H}_1^{(l)}, \ldots, f_{d^*} \in \mathcal{H}_{d^*}^{(l)} \right\}.$$

If we choose this function space in our estimate (1), we get the following result.

Theorem 1: Let (X, Y), (X_1, Y_1), (X_2, Y_2), ...be independent and identically distributed random variables with values in $\mathbb{R}^d \times \mathbb{R}$ such that

$$\mathbf{E} \exp\left(c_1 \cdot Y^2\right) < \infty$$

for some constant $c_1 > 0$. Let m be the corresponding regression function and assume that m satisfies a hierarchical interaction model of order d^* and level $l \in \mathbb{N}_0$, which is (p, C)–smooth according to Definition 3 for some $p \in \mathbb{N}$ and $C > 0$. Furthermore assume that $supp(X)$ is bounded.

Let m_n be the least squares estimate defined by (1), where the function space is chosen as above using tensor product spline functions and where the construction is done accordingly to the hierarchical interaction model for m with parameters

$$K = K_n = \left\lceil n^{1/(2p+d^*)} \right\rceil, \alpha = \alpha_n = \log n, \beta = \beta_n = \log n$$
$$\text{and } \gamma = \gamma_n = \log n$$

and degree $M > p - 1$. Then we have for n sufficiently large

$$\mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \leq c_2 \cdot \log^{\max\{3,2p\}}(n) \cdot n^{-2p/(2p+d^*)}.$$

Remark 2: In the definition of the estimate in Theorem 1 we have to choose parameters depending on the smoothness and the structure of the assumed hierarchical interaction model, which is not possible in an application since there the smoothness of the regression function will be usually unknown. But there are standard data-driven methods to choose the parameters of a regression estimate, e.g., splitting of the sample (cf., e.g., Györfi *et al.* [9, Ch. 7]). If we apply splitting of the sample, then the result of Theorem 1 can be shown also for an estimate whose definition does not depend on the smoothness of the regression function.

IV. Estimates Based on Neural Networks

In this section we assume that the function space $\mathcal{H}_n$ in the definition of our least squares estimate (1) consists of multilayer feedforward neural networks. The starting point in defining such neural networks is the choice of a so-called activation function $\sigma : \mathbb{R} \to [0, 1]$. Usually one uses here so-called squashing activation functions which are defined as functions $\sigma : \mathbb{R} \to [0, 1]$ which are nondecreasing and satisfy

$$\lim_{x \to -\infty} \sigma(x) = 0 \quad \text{and} \quad \lim_{x \to \infty} \sigma(x) = 1.$$

Examples of activation functions which are squashing functions include the sigmoidal squasher

$$\sigma(x) = \frac{1}{1 + \exp(-x)}$$

or the Gaussian squasher

$$\sigma(x) = \frac{1}{\sqrt{2 \cdot \pi}} \int_{-\infty}^{x} \exp(-u^2/2)\, du.$$

Multilayer feedforward neural networks with sigmoidal functions can be defined recursively as follows: A multilayer feedforward neural network with l hidden layers, K_1, ..., $K_l \in \mathbb{N}$ neurons in the first, second, ..., l-th layer, respectively, and sigmoidal function σ is a real-valued function defined on $\mathbb{R}^d$ of the form

$$f(x) = \sum_{i=1}^{K_l} c_i^{(l)} \cdot f_i^{(l)}(x) + c_0^{(l)}, \tag{2}$$

for some $c_0^{(l)}, \ldots, c_{K_l}^{(l)} \in \mathbb{R}$ and for $f_i^{(l)}$'s recursively defined by

$$f_i^{(r)}(x) = \sigma\left(\sum_{j=1}^{K_{r-1}} c_{i,j}^{(r-1)} \cdot f_j^{(r-1)}(x) + c_{i,0}^{(r-1)}\right) \tag{3}$$

for some $c_{i,0}^{(r-1)}, \ldots, c_{i,K_{r-1}}^{(r-1)} \in \mathbb{R}$ and

$$f_i^{(1)}(x) = \sigma\left(\sum_{j=1}^{d} c_{i,j}^{(0)} \cdot x^{(i)} + c_{i,0}^{(0)}\right) \tag{4}$$

for some $c_{i,0}^{(0)}, \ldots, c_{i,d}^{(0)} \in \mathbb{R}$.

For applications of neural networks to nonlinear function estimation, classification and learning we refer the reader to the monographs Hertz *et al.* [15], Devroye *et al.* [6], Anthony and Bartlett [1], Györfi *et al.* [9], Hastie *et al.* [13], Haykin [14], and Ripley [24].

Consistency of nonparametric regression estimates using neural networks has been studied by Lugosi and Zeger [20] and Mielniczuk and Tyrcha [23]. The rate of convergence of neural network regression estimates with one hidden layer has been analyzed by Barron [4], [5] and McCaffrey and Gallant [21], and in connection with feedforward neural network with two hidden layers in Kohler and Krzyżak [18].

Our choice of the set of neural networks suitable for estimation of generalized hierarchical interaction models is motivated by the following approximation result presented in Mhaskar [22]: Let $m : \mathbb{R}^d \to \mathbb{R}$ be a (p, C)–smooth function, where $0 < p \leq 1$, let $N \in \mathbb{N}$ and let A be a compact subset of $\mathbb{R}^d$. Then there exists a neural network

$$t(x) = \sum_{i=1}^{N^d} c_i \cdot \sigma\left(\sum_{j=1}^{d} b_{i,j} \cdot \sigma\left(\sum_{k=1}^{d} a_{i,j,k} \cdot x^{(k)} + a_{i,j,0}\right) + b_{i,0}\right) + c_0$$

with two hidden layers such that

$$|t(x) - m(x)| \leq c_3 \cdot C \cdot \frac{1}{N^p}$$

for "nearly" all $x \in A$ (see Lemma 6 below for details).

Now assume that m satisfies a generalized hierarchical interaction model of order d^* and level 0, which is (p, C)–smooth, i.e.,

$$m(x) = f(a_1^T x, \ldots, a_{d^*}^T x) \quad \text{for all } x \in \mathbb{R}^d$$

for some $a_1, \ldots, a_{d^*} \in \mathbb{R}^{d^*}$ and some (p, C)-smooth function $f : \mathbb{R}^{d^*} \to \mathbb{R}$. Approximating f by the above feedforward neural network with two hidden layers defined on $\mathbb{R}^{d^*}$ we

see that we can approximate m by the following feedforward neural network with two hidden layers defined on $\mathbb{R}^d$:

$$f(x) = \sum_{i=1}^{N^{d^*}} c_i \cdot \sigma\left(\sum_{j=1}^{d^*} b_{i,j} \cdot \sigma\left(\sum_{k=1}^{d} a_{i,j,k} \cdot x^{(k)} + a_{i,j,0}\right) + b_{i,0}\right) + c_0 \quad (x \in \mathbb{R}^d). \tag{5}$$

Here in the first and in the second hidden layer we are using $d^* \cdot N^{d^*}$ and N^{d^*} neurons, respectively. However, the neural network has only

$$N^{d^*} + 1 + N^{d^*} \cdot (d^* + 1) + N^{d^*} \cdot d^* \cdot (d + 1) = N^{d^*} \cdot (d^* \cdot d + 2 \cdot d^* + 1) + 1 \tag{6}$$

weights. This is due to the fact, that the two hidden layers of the neural network are not fully connected. Instead, each neuron in the second hidden layer is connected with d^* neurons in the first hidden layer, and this is done in such a way that each neuron in the first hidden layer is connected with exactly one neuron in the second hidden layer.

For $N \in \mathbb{N}$, $d \in \mathbb{N}$, $d^* \in \{1, \ldots, d\}$ and $\alpha > 0$ we denote the sets of all functions $f : \mathbb{R}^d \to \mathbb{R}$ which satisfy (5) for some $a_{i,j,k}, b_{i,j}, c_i \in \mathbb{R}$, where

$$|a_{i,j,k}| \le \alpha, \quad |b_{i,j}| \le \alpha \quad \text{and } |c_i| \le \alpha$$

for all $i \in \{0, 1, \ldots, N^{d^*}\}$, $j \in \{0, 1, \ldots, d^*\}$ and $k \in \{0, 1, \ldots, d\}$, by $\mathcal{F}_{N,d^*,d,\alpha}^{(neural\ networks)}$. Motivated by the definition of a generalized hierarchical interaction model, we define so–called spaces of hierarchical neural networks with parameters K, N, d^*, d and level l as follows. In case $l = 0$ we set

$$\mathcal{H}^{(0)} = \mathcal{F}_{N,d^*,d,\alpha}^{(neural\ networks)}.$$

And for $l > 0$ we define

$$\mathcal{H}^{(l)} = \Bigg\{ h : \mathbb{R}^d \to \mathbb{R} \quad : \quad h(x) = \sum_{k=1}^{K} g_k\left(f_{1,k}(x), \ldots, f_{d^*,k}(x)\right) \quad (x \in \mathbb{R}^d) \quad \text{for some } g_k \in \mathcal{F}_{N,d^*,d^*,\alpha}^{(neural\ networks)} \text{ and } f_{j,k} \in \mathcal{H}^{(l-1)} \Bigg\}.$$

The class $\mathcal{H}^{(0)}$ is a set neural networks with two hidden layers and number of weights given by (6). From this one can conclude recursively that for $l > 0$ the class $\mathcal{H}^{(l)}$ is a set neural networks with $2 \cdot l$ hidden layers, where the weights can be parameterized by

$$(K + 1)^l \cdot \left(N^{d^*} \cdot (d + 1)^2 + 1\right)$$

many parameters (there are in fact much more weights in the neural networks, however, the are related to each other (in the sense that they are products of weights $a_{i,j,k}$ of the network at level $2r + 1$ and of weights c_i of the network at level $2r$) and can therefore be parameterized by the above number of parameters).

Next we choose in our least squares estimate (1) the set $\mathcal{H}_n$ as the set $\mathcal{H}^{(l)}$, with parameters $K = K_{max}$, $N = K_n$, d^*, d and level l, where d^* and l are the values from the definition of the generalized hierarchical interaction model for m. In the next result we need to truncate our regression estimate. We define the truncation operator T_L as follows

$$T_L u = \begin{cases} u & \text{if } |u| \le L, \\ L \cdot \operatorname{sign}(u) & \text{otherwise,} \end{cases}$$

Then the following result holds.

Theorem 2: Let (X, Y), (X_1, Y_1), (X_2, Y_2), ...be independent and identically distributed random variables with values in $\mathbb{R}^d \times \mathbb{R}$ such that

$$\mathbf{E} \exp\left(c_1 \cdot Y^2\right) < \infty$$

for some constant $c_1 > 0$ and such that $supp(X)$ is bounded. Let m be the corresponding regression function and assume that m satisfies a generalized hierarchical interaction model of order d^* which is (p, C)–smooth according to Definition 3 for some $0 < p \le 1$ and $C > 0$ and where all functions occurring in Definition 2 b) are Lipschitz continuous. Let K_{max} be the maximal number of summands in the different levels in Definition 2 b).

Let m_n be the least squares estimate defined by (1) with $\mathcal{H}_n$ defined as above with

$$K_n = \left\lceil \left(\frac{n}{\log(n)}\right)^{1/(2p+d^*)} \right\rceil \quad \text{and} \quad \alpha_n = n^4.$$

Assume that the sigmoidal function $\sigma : \mathbb{R} \to [0, 1]$ is a Lipschitz continuous squashing function which satisfies

$$|\sigma(y) - 1| \le \frac{1}{y} \quad \text{if } y > 0 \quad \text{and } |\sigma(y)| \le \frac{1}{|y|} \text{ if } y < 0.$$

Then

$$\mathbf{E} \int |T_{const \cdot \log n} m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \le c_4 \cdot \log^3(n) \cdot n^{-2p/(2p+d^*)}.$$

Remark 3: The class of (p, C)–smooth generalized hierarchical interaction models of order d^*, where all functions occurring in Definition 2 b) are Lipschitz continuous, contains all (p, C)–smooth functions which depend on at most d^* of its input components, since in the definition of generalized hierarchical interaction models all functions occurring in Definition 2 might be chosen as projections. Consequently the rate of convergence in Theorem 2 is optimal up to some logarithmic factor according to Stone [25].

Remark 4: As in Remark 2 the parameters of our neural network estimate can be chosen in a data-dependent way by splitting of the sample.

V. Proofs

A. A General Result on Least Squares Estimates

The estimates in Theorems 1 and 2 are least squares estimates. The L_2 error of such estimates depends on the approximation properties and the complexity of the used

functions spaces. The latter one can be measured by so-called covering numbers, which we introduce next.

Definition 4: Let $\epsilon > 0$, let $\mathcal{G}$ be a set of functions $g : \mathbb{R}^d \to \mathbb{R}$, let $x_1, \ldots, x_n \in \mathbb{R}^d$ and set $x_1^n = (x_1, \ldots, x_n)$.

a) *Every finite collection of functions $g_1, \ldots, g_N : \mathbb{R}^d \to \mathbb{R}$ with the property that for every $g \in \mathcal{G}$ there exists a $j = j(g) \in \{1, \ldots, N\}$ such that*

$$\|g - g_j\|_\infty < \epsilon$$

is called a supremum norm ϵ–cover of $\mathcal{G}$. The size of the smallest supremum norm ϵ–cover of $\mathcal{G}$ is called supremum norm ϵ–covering number of $\mathcal{G}$ and is denoted by $\mathcal{N}_\infty(\epsilon, \mathcal{G})$. Here we set $\mathcal{N}_\infty(\epsilon, \mathcal{G}) = \infty$ in case that there exists no finite supremum norm ϵ–cover of $\mathcal{G}$.

b) *Every finite collection of functions $g_1, \ldots, g_N : \mathbb{R}^d \to \mathbb{R}$ with the property that for every $g \in \mathcal{G}$ there exists a $j = j(g) \in \{1, \ldots, N\}$ such that*

$$\frac{1}{n}\sum_{i=1}^{n} |g(x_i) - g_j(x_i)| < \epsilon$$

is called a L_1 norm ϵ–cover of $\mathcal{G}$ on x_1^n. The size of the smallest L_1 norm ϵ–cover of $\mathcal{G}$ on x_1^n is called L_1 norm ϵ–covering number of $\mathcal{G}$ on x_1^n and is denoted by $\mathcal{N}_1(\epsilon, \mathcal{G}, x_1^n)$. Here we set $\mathcal{N}_1(\epsilon, \mathcal{G}, x_1^n) = \infty$ in case that there exists no finite L_1 norm ϵ–cover of $\mathcal{G}$ on x_1^n.

Using the notion of covering numbers we can formulate the following general result on the least squares estimates.

Lemma 1: Let (X, Y), (X_1, Y_1), (X_2, Y_2), ...be independent and identically distributed random variables with values in $\mathbb{R}^d \times \mathbb{R}$ such that

$$\mathbf{E}\exp\left(c_1 \cdot Y^2\right) < \infty$$

for some constant $c_1 > 0$. Let m be the corresponding regression function and let m_n be the least squares estimate defined by (1). Assume that the function space $\mathcal{H}_n$ consists of functions bounded in absolute value by $c_5 \cdot \log(n)$ and that its covering number satisfies

$$\sup_{x_1,\ldots,x_n \in \mathbb{R}^d} \mathcal{N}_1\left(\frac{1}{n}, \mathcal{H}_n, x_1^n\right) \le \mathcal{N}_1\left(\frac{1}{n}, \mathcal{H}_n\right)$$

for some $\mathcal{N}_1\left(\frac{1}{n}, \mathcal{H}_n\right) \ge 3$. Then

$$\begin{aligned}&\mathbf{E}\int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx)\\ &\le c_5 \cdot \log(n)^2 \cdot \frac{\log\left(\mathcal{N}_1\left(\frac{1}{n}, \mathcal{H}_n\right)\right)}{n}\\ &\quad + 2 \inf_{h \in \mathcal{H}_n} \int |h(x) - m(x)|^2 \mathbf{P}_X(dx).\end{aligned}$$

Proof: The result is a consequence of the standard error bounds on least squares estimates derived by using results from the empirical process theory, cf., e.g., proof of Györfi *et al.* [9, Th. 11.5] and proof of Bagirov *et al.* [3, Th. 1]. □

Remark 5: Lemma 1 also holds whenever a function space consists of unbounded functions, if one truncates the least squares estimate at level $\pm c_s \times \log(n)$.

B. Proof of Theorem 1

In the proof of Theorem 1 we will apply Lemma 1. In order to bound the covering number and the approximation error (i.e., $\inf_{h \in \mathcal{H}_n} \int |h(x) - m(x)|^2 \mathbf{P}_X(dx)$), we will need the following auxiliary results.

Lemma 2: Let $g, \bar{g} : \mathbb{R}^d \to \mathbb{R}$, $f_1, \bar{f}_1, \ldots, f_d, \bar{f}_d : \mathbb{R}^d \to \mathbb{R}$ and define m and $\bar{m}$ by

$$m(x) = g(f_1(x), \ldots, f_d(x)) \quad (x \in \mathbb{R}^d)$$

and

$$\bar{m}(x) = \bar{g}(\bar{f}_1(x), \ldots, \bar{f}_d(x)) \quad (x \in \mathbb{R}^d).$$

If g is Lipschitz continuous with Lipschitz constant $C > 0$, then we have for any $x \in \mathbb{R}^d$

$$|m(x) - \bar{m}(x)| \le C \cdot \sqrt{d} \cdot \sum_{j=1}^{d} |f_j(x) - \bar{f}_j(x)| + \|g - \bar{g}\|_\infty.$$

Proof: The result follows from the triangle inequality, the Lipschitz continuity of g and a bound on the L_2 norm by the L_1 norm:

$$\begin{aligned}&|m(x) - \bar{m}(x)|\\ &\le \left|g(f_1(x), \ldots, f_d(x)) - g(\bar{f}_1(x), \ldots, \bar{f}_d(x))\right|\\ &\quad + \left|g(\bar{f}_1(x), \ldots, \bar{f}_d(x)) - \bar{g}(\bar{f}_1(x), \ldots, \bar{f}_d(x))\right|\\ &\le C \cdot \left\|\left(f_1(x) - \bar{f}_1(x), \ldots, f_d(x) - \bar{f}_d(x)\right)^T\right\| + \|g - \bar{g}\|_\infty\\ &\le C \cdot \sqrt{d} \cdot \sum_{j=1}^{d} |f_j(x) - \bar{f}_j(x)| + \|g - \bar{g}\|_\infty.\end{aligned}$$

□

Let $\mathcal{G}$, $\mathcal{F}_1$, ..., $\mathcal{F}_d$ be sets of functions $f : \mathbb{R}^d \to \mathbb{R}$ and define

$$\begin{aligned}\mathcal{H} &= \mathcal{G}(\mathcal{F}_1, \ldots, \mathcal{F}_d)\\ &= \Big\{h : \mathbb{R}^d \to \mathbb{R} \ : h(x) = g(f_1(x), \ldots, f_d(x)) \quad (x \in \mathbb{R}^d)\\ &\qquad \text{for some } g \in \mathcal{G}, f_1 \in \mathcal{F}_1, \ldots, f_d \in \mathcal{F}_d\Big\}.\end{aligned}$$

Lemma 3: Let $\mathcal{G}$, $\mathcal{F}_1$, ..., $\mathcal{F}_d$ be sets of functions $f : \mathbb{R}^d \to \mathbb{R}$ and define $\mathcal{H} = \mathcal{G}(\mathcal{F}_1, \ldots, \mathcal{F}_d)$ as above. If the functions in $\mathcal{G}$ are Lipschitz continuous with Lipschitz constant $C > 0$, then we have for any $x_1^n \in (\mathbb{R}^d)^n$ and any $\epsilon > 0$:

$$\mathcal{N}_1(\epsilon, \mathcal{H}, x_1^n) \le \mathcal{N}_\infty\left(\frac{\epsilon}{2}, \mathcal{G}\right) \cdot \prod_{j=1}^{d} \mathcal{N}_1\left(\frac{\epsilon}{2 \cdot \sqrt{d} \cdot d \cdot C}, \mathcal{F}_j, x_1^n\right).$$

Proof: Follows directly from Lemma 2. □

Lemma 4: Let $\mathcal{G}$, $\mathcal{F}_1$, ..., $\mathcal{F}_d$ be sets of functions $\mathbb{R}^d \to \mathbb{R}$ and define $\mathcal{H} = \mathcal{G}(\mathcal{F}_1, \ldots, \mathcal{F}_d)$ as above. Let $g, f_1, \ldots, f_d : \mathbb{R}^d \to \mathbb{R}$ and define m by

$$m(x) = g(f_1(x), \ldots, f_d(x)) \quad (x \in \mathbb{R}^d).$$

If g is Lipschitz continuous with Lipschitz constant $C > 0$, then

$$\inf_{h \in \mathcal{H}} \|m - h\|_\infty \le \sqrt{d} \cdot C \cdot \sum_{j=1}^{d} \inf_{\bar{f} \in \mathcal{F}_j} \|f_j - \bar{f}\|_\infty + \inf_{\bar{g} \in \mathcal{G}} \|g - \bar{g}\|_\infty.$$

Proof: Follows directly from Lemma 2. □

The following lemma describes a bound on the approximation error of the tensor product spline spaces introduced in Section 3.

Lemma 5: Let $p \in \mathbb{N}$ and $C > 0$ and let $m : \mathbb{R}^d \to \mathbb{R}$ be a (p, C)–smooth function with compact support. Let $M > p-1$, $K \in \mathbb{N}$ and set

$$\alpha = \alpha_n = \log n, \quad \beta = \beta_n = \log n \quad \text{and} \quad \gamma = \gamma_n = \log n$$

and define the tensor product spline space $\mathcal{S}_{K,M,\alpha,\beta,\gamma,d}$ as in Section 3. Then we have for n sufficiently large

$$\inf_{f \in \mathcal{S}_{K,M,\alpha,\beta,\gamma,d}} \|m - f\|_\infty \le c_6 \cdot \left(\frac{\log(n)}{K}\right)^p.$$

Proof: Follows from Lemma 15.2, Theorem 15.1 and the proof of Györfi *et al.* [9, Th. 15.2]. Here we use the fact that the coefficients of the spline approximand constructed in the proof of Györfi *et al.* [9, Th. 15.2] satisfy

$$|a_{\mathbf{k}}| \le \log n \quad \text{and} \quad |a_{\mathbf{k}} - a_{\mathbf{k}-e_i}| \le \frac{(\log n)^2}{\sqrt{d} \cdot K}$$

for n sufficiently large, since $a_{\mathbf{k}}$ is a linear combination of point evaluations of the bounded function m and since $a_{\mathbf{k}} - a_{\mathbf{k}-e_i}$ is a linear combinations of differences of point evaluations of the Lipschitz continuous function m at points which have a supremum norm distance less than or equal to $(2M+2) \cdot \beta / K$. □

Proof of Theorem 1: An easy discretization of the (bounded) coefficients in the definition of the spline space $\mathcal{S}_{K,M,\alpha,\beta,\gamma,d^*}$ together with Györfi *et al.* [9, Lemma 15.2] shows that

$$\mathcal{N}_\infty(\epsilon, \mathcal{S}_{K,M,\alpha,\beta,\gamma,d^*}) \le \left(\frac{2 \cdot \log n}{\epsilon}\right)^{(2 \cdot K + M)^{d^*}}.$$

From this we get by a (w.r.t. l) recursive application of Lemma 3

$$\sup_{x_1,\ldots,x_n \in \mathbb{R}^d} \mathcal{N}_1\left(\frac{1}{n}, \mathcal{H}_n^{(l)}, x_1^n\right) \le n^{c_7 \cdot K^{d^*}}$$

for n sufficiently large (for some constant c_7 which depends on l). Furthermore, recursive application of Lemma 4 again together with Lemma 5 and the Lipschitz smoothness of all functions g occuring in Definition 1 b) of the hierarchical interaction model from m (which follows from the (p, C)–smoothness of the model and $p \ge 1$) implies

$$\inf_{h \in \mathcal{H}_n} \int |h(x) - m(x)|^2 \mathbf{P}_X(dx) \le \inf_{h \in \mathcal{H}_n} \|h(x) - m(x)\|^2_{\infty, supp(X)} \le c_8 \cdot \left(\frac{\log(n)}{K}\right)^{2p}$$

for n sufficiently large. Using these two bounds we get the assertion by an application of Lemma 1 and the definition of K. □

C. Proof of Theorem 2

In the proof we will use the following auxiliary results.

Lemma 6: Let $m : \mathbb{R}^d \to \mathbb{R}$ be a (p, C)–smooth function, where $0 < p \le 1$, let $N \in \mathbb{N}$, let $A \supseteq [0,1]^d$ be a compact subset of $\mathbb{R}^d$, let $\eta \in (0, 1]$ and let ν be a probability measure on $\mathbb{R}^d$. Let $\sigma : \mathbb{R} \to [0, 1]$ be a squashing function. Then there exists a neural network

$$t(x) = \sum_{i=1}^{N^d} c_i \cdot \sigma\left(\sum_{j=1}^{d} b_{i,j} \cdot \sigma\left(\sum_{k=1}^{d} a_{i,j,k} \cdot x^{(k)} + a_{i,j,0}\right) + b_{0,j}\right) + c_0$$

with two hidden layers such that outside of a set of ν-measure less than or equal to η we have for all $x \in A$

$$|t(x) - m(x)| \le c_9 \cdot \frac{1}{N^p}.$$

In case that σ satisfies

$$|\sigma(y) - 1| \le \frac{1}{y} \quad \text{if } y > 0 \quad \text{and } |\sigma(y)| \le \frac{1}{|y|} \quad \text{if } y < 0$$

the weights in the neural network above can be chosen such that

$$|c_i| \le 2^{d+1} \cdot \|m\|_\infty, \quad |b_{i,j}| \le 4 \cdot d \cdot N^d$$
$$\text{and } |a_{i,j,k}| \le 24 \cdot d^2 \cdot (\max_{z \in A} \|z\|_\infty + 1) \cdot \frac{N}{\eta}$$

($i \in \{1, \ldots, N^d\}$, $j, k \in \{1, \ldots, d\}$).

Proof: The result can be proven by modifying the proof of Mhaskar [22, Th. 3.4]. For the sake of completeness we present a complete proof of this result in the Appendix. □

Lemma 7: Let $\sigma : \mathbb{R} \to [0, 1]$ be a sigmoidal function which is Lipschitz continuous with Lipschitz constant $C \ge 1$. Define f, $\bar{f} : \mathbb{R}^d \to \mathbb{R}$ recursively by

$$f(x) = \sum_{i=1}^{K_l} c_i^{(l)} \cdot f_i^{(l)}(x) + c_0^{(l)}$$

and

$$\bar{f}(x) = \sum_{i=1}^{K_l} \bar{c}_i^{(l)} \cdot \bar{f}_i^{(l)}(x) + \bar{c}_0^{(l)}$$

for some $c_0^{(l)}, \bar{c}_0^{(l)}, \ldots, c_{K_l}^{(l)}, \bar{c}_{K_l}^{(l)} \in \mathbb{R}$ and for $f_i^{(l)}$, $\bar{f}_i^{(l)}$'s recursively defined by

$$f_i^{(r)}(x) = \sigma\left(\sum_{j=1}^{K_{r-1}} c_{i,j}^{(r-1)} \cdot f_j^{(r-1)}(x) + c_{i,0}^{(r-1)}\right)$$

and

$$\bar{f}_i^{(r)}(x) = \sigma\left(\sum_{j=1}^{K_{r-1}} \bar{c}_{i,j}^{(r-1)} \cdot \bar{f}_j^{(r-1)}(x) + \bar{c}_{i,0}^{(r-1)}\right)$$

for some $c_{i,0}^{(r-1)}, \bar{c}_{i,0}^{(r-1)}, \ldots, c_{i,K_{r-1}}^{(r-1)}, \bar{c}_{i,K_{r-1}}^{(r-1)} \in \mathbb{R}$ ($r \in \{2, \ldots, l\}$, $i \in \{1, \ldots, K_r\}$) and

$$f_i^{(1)}(x) = \sigma\left(\sum_{j=1}^{d} c_{i,j}^{(0)} \cdot x^{(j)} + c_{i,0}^{(0)}\right)$$

and

$$\bar{f}_i^{(1)}(x) = \sigma\left(\sum_{j=1}^{d} \bar{c}_{i,j}^{(0)} \cdot x^{(j)} + \bar{c}_{i,0}^{(0)}\right)$$

for some $c_{i,0}^{(0)}, \ldots, c_{i,d}^{(0)}, \bar{c}_{i,0}^{(0)}, \ldots, \bar{c}_{i,d}^{(0)} \in \mathbb{R}$ *(*$i \in \{1, \ldots, K_1\}$*). Then*

$$
\begin{aligned}
&|f(x) - \bar{f}(x)| \\
&\leq \max\{\|x\|_\infty, 1\} \cdot (d+1) \cdot (l+1) \\
&\quad \cdot \left(1 + \max_{r=0,\ldots,l, i=1,\ldots,K_{r+1}, j=1,\ldots,K_r} |c_{i,j}^{(r)}|\right)^{l+1} \\
&\quad \cdot C^l \cdot \prod_{r=0}^{l} (K_r + 1) \cdot \max_{r=0,\ldots,l, i=1,\ldots,K_{r+1}, j=0,\ldots,K_r} |c_{i,j}^{(r)} - \bar{c}_{i,j}^{(r)}|
\end{aligned}
$$

for any $x \in \mathbb{R}^d$*, where we have set* $K_0 = d$*,* $K_{l+1} = 1$ *and* $c_{1,i}^{(l)} = c_i^{(l)}$.

Proof: By the triangle inequality and $\|\sigma\|_\infty \leq 1$ we get

$$
\begin{aligned}
&|f(x) - \bar{f}(x)| \\
&\quad \leq \sum_{i=1}^{K_l} |c_i^{(l)}| \cdot |f_i^{(l)}(x) - \bar{f}_i^{(l)}(x)| \\
&\qquad + \sum_{i=1}^{K_l} |c_i^{(l)} - \bar{c}_i^{(l)}| \cdot |\bar{f}_i^{(l)}(x)| + |c_0^{(l)} - \bar{c}_0^{(l)}| \\
&\quad \leq K_l \cdot \max_{i=1,\ldots,K_l} |c_i^{(l)}| \cdot \max_{i=1,\ldots,K_l} |f_i^{(l)}(x) - \bar{f}_i^{(l)}(x)| \\
&\qquad + (K_l + 1) \cdot \max_{i=0,\ldots,K_l} |c_i^{(l)} - \bar{c}_i^{(l)}|.
\end{aligned}
$$

Using the Lipschitz continuity of σ and again the triangle inequality we get furthermore

$$
\begin{aligned}
&|f_i^{(r)}(x) - \bar{f}_i^{(r)}(x)| \\
&\quad \leq C \cdot |\sum_{j=1}^{K_{r-1}} c_{i,j}^{(r-1)} \cdot f_j^{(r-1)}(x) + c_{i,0}^{(r-1)} \\
&\qquad - \sum_{j=1}^{K_{r-1}} \bar{c}_{i,j}^{(r-1)} \cdot \bar{f}_j^{(r-1)}(x) - \bar{c}_{i,0}^{(r-1)}| \\
&\quad \leq C \cdot K_{r-1} \cdot \max_{j=1,\ldots,K_{r-1}} |c_{i,j}^{(r-1)}| \\
&\qquad \cdot \max_{j=1,\ldots,K_{r-1}} |f_j^{(r-1)}(x) - \bar{f}_j^{(r-1)}(x)| \\
&\qquad + C \cdot (K_{r-1} + 1) \cdot \max_{j=0,\ldots,K_{r-1}} |c_{i,j}^{(r-1)} - \bar{c}_{i,j}^{(r-1)}|.
\end{aligned}
$$

Finally, in the same way we see

$$
\begin{aligned}
&|f_i^{(1)}(x) - \bar{f}_i^{(1)}(x)| \\
&\quad \leq C \cdot \max\{\|x\|_\infty, 1\} \cdot (d+1) \cdot \max_{j=0,\ldots,d} |c_{i,j}^{(0)} - \bar{c}_{i,j}^{(0)}|,
\end{aligned}
$$

which implies the assertion. □

Lemma 8: Let $\sigma : \mathbb{R} \to [0,1]$ *be a sigmoidal function which is Lipschitz continuous with Lipschitz constant* $C \geq 1$ *and let* A *be a compact subset of* $\mathbb{R}^d$*. Then for any* $l \in \mathbb{N}$*,* $x_1^n \in A^n$*,* $d^* \in \{1, \ldots, d\}$*,* $K_n \geq 2$ *and* $\alpha_n \geq 2$ *we have*

$$
\mathcal{N}_1(\epsilon, \mathcal{H}^{(l)}, x_1^n) \leq c_{10} \cdot \left(\frac{\alpha_n \cdot K_n}{\epsilon}\right)^{c_{10} \cdot K_n^{d^*}}
$$

for some constant c_{10}*, which depends on* l*,* d *and* d^**.*

Proof: The neural networks in $\mathcal{H}^{(l)}$ can be parameterized using

$$
(K_{max} + 1)^l \cdot \left(N^{d^*} \cdot (d+1)^2 + 1\right)
$$

parameters. Discretizing them using a grid of size $\delta > 0$ results in a set of functions of size

$$
\left(\frac{2 \cdot \alpha_n}{\delta}\right)^{c_{11} \cdot K_n^{d^*}},
$$

which, according to Lemma 7 has the property that for each $f \in \mathcal{H}^{(l)}$ there exists a function $\bar{f}$ in this set satisfying

$$
|f(x) - \bar{f}(x)| \leq c_{12} \cdot (K_n \cdot \alpha_n)^{c_{12}} \cdot \delta.
$$

Here we have used the fact that some of the weights of the neural network are products of the above parameters, and that for such products we have

$$
\begin{aligned}
\left|a \cdot b - \bar{a} \cdot \bar{b}\right| &\leq |a - \bar{a}| \cdot |b| + \left|b - \bar{b}\right| \cdot |\bar{a}| \\
&\leq 2 \cdot \max\{|\bar{a}|, |b|\} \cdot \max\{|a - \bar{a}|, \left|b - \bar{b}\right|\}.
\end{aligned}
$$

The result follows by setting

$$
\delta = \frac{\epsilon}{c_{12} \cdot (K_n \cdot \alpha_n)^{c_{12}}}.
$$

□

Proof of Theorem 2: Repeated application of Lemma 4, which is possible because of the Lipschitz continuity of the functions occuring in Definition 2 b), together with Lemma 6 imply

$$
\inf_{h \in \mathcal{H}_n} \int |h(x) - m(x)|^2 \mathbf{P}_X(dx) \leq c_{13} \cdot \left(\frac{1}{K_n}\right)^{2p}
$$

for n sufficiently large. Here we use that for a generalized hierarchical interaction model the bound on the approximation error in Lemma 6 holds simultaneously for all occuring functions outside of an event of $\mathbf{P}_X$–measure at most c_{14}/n^2. And on this event the integrand in the above integral is bounded in absolute value by $c_{15} \cdot K_n^{d^*} \leq c_{16} \cdot n$.

Furthermore, by Lemma 8 we can bound the covering number by

$$
\mathcal{N}_1(\frac{1}{n}, \mathcal{H}_n, x_1^n) \leq c_{14} \cdot n^{c_{15} \cdot K_n^{d^*}}
$$

for any $x_1, \ldots, x_n \in supp(X)$. Using these two bounds we get the assertion by an application of Lemma 1, Remark 4 and the definition of K_n. □

VI. Conclusion

In this paper we investigate d-dimensional regression functions with generalized hierarchical interaction type structures. These functions are computed in several layers, such that in each layer a function of at most d* inputs is computed. Our models encompass additive regression and projection pursuit models studied earlier in the literature. We investigated convergence of two nonparametric regression estimates of hierarchical interaction models: one based on polynomial splines and the other on neural networks. We demonstrated that the rates of convergence of the estimates depend on d* rather than on d thus circumventing the curse of dimensionality.

APPENDIX

In the proof we will use Mhaskar [22, Proposition 3.8], which we reformulate here (in a slightly different form) as Lemma 9.

Lemma 9: Let $K \subseteq \mathbb{R}^d$ be a polytope bounded by hyperplanes $v_i \cdot x + d_i \geq 0$ $(i = 1, \ldots, L)$, where $v_1, \ldots, v_L \in \mathbb{R}^d$ and $d_1, \ldots, d_L \in \mathbb{R}$. For $\delta > 0$ set

$$K_\delta^0 := \left\{ x \in \mathbb{R}^d \, : \, v_i \cdot x + d_i \geq \delta \ \text{ for all } i \in \{1, \ldots, L\} \right\}$$

and

$$K_\delta^c := \left\{ x \in \mathbb{R}^d : v_i \cdot x + d_i \leq -\delta \ \text{ for some } i \in \{1, \ldots, L\} \right\}.$$

Let $\sigma : \mathbb{R} \to [0, 1]$ be a squashing function. Let $\epsilon, \delta \in (0, 1]$ be arbitrary. Then there exists a neural network of the form

$$f(x) = \sigma \left(\sum_{i=1}^{L} b_i \cdot \sigma \left(\sum_{j=1}^{d} c_{i,j} \cdot x^{(i)} + c_{i,0} \right) + b_0 \right)$$

satisfying

$$\begin{aligned} &|f(x)| \leq 1 \quad \text{for} \quad x \in \mathbb{R}^d, \\ &|f(x) - 1| \leq \epsilon \quad \text{for} \quad x \in K_\delta^0, \\ &|f(x)| \leq \epsilon \quad \text{for} \quad x \in K_\delta^c. \end{aligned} \tag{7}$$

In case that the squashing function satisfies

$$|\sigma(y) - 1| \leq \frac{1}{y} \ \text{ if } \ y > 0 \ \text{ and } \ |\sigma(y)| \leq \frac{1}{|y|} \ \text{ if } \ y < 0,$$

the weights above can be chosen such that

$$|b_i| \leq \frac{4L}{\epsilon},$$

and

$$|c_{i,j}| \leq \frac{4 \cdot L}{\delta} \cdot \max\{\|v_1\|_\infty, |d_1|, \ldots, \|v_L\|_\infty, |d_L|\}$$

$(i = 0, \ldots, L, j = 0, \ldots, d)$.

Proof: Follows from the proof of Mhaskar [22, Proposition 3.8]. □

Proof of Lemma 6: W.l.o.g we assume that A is a cube. We partition this cube into N^d equivolume cubes of side length c_{16}/N (where $c_{16} \geq 1$ since $[0, 1]^d \subseteq A$). Approximating m by a piecewise constant approximand with respect to this partition yields (since m is (p, C)–smooth) a function S satisfying

$$\|S - m\|_{\infty, A} \leq c_{16} \cdot N^{-p}. \tag{8}$$

S can be expressed in the form

$$S(x) = m(x_0) + \sum_{j \in \{1, \ldots, N\}^d} d_j \cdot \prod_{i=1}^{d} (x^{(i)} - x_j^{(i)})_+^0,$$

where x_j are the corners of the rectangles comprising the above partition and d_j are constants satisfying

$$|d_j| \leq c_{17} \cdot N^{-p}$$

(constructed by using differences of function values of m at the corners of the above partition) and $x_+ = \max\{x, 0\}$. Let K_j be the polytope defined by $x^{(i)} - x_j^{(i)} \geq 0$ $(i = 1, \ldots, d)$. Set $\epsilon = N^{-d}$, $\delta = \eta/(6 \cdot d \cdot N)$ and apply Lemma 9 for each K_j (i.e., with $L = d$, $v_i = \mathbf{e}_i$ and $b_i = -x_j^{(i)}$, where $\mathbf{e}_i$ denotes the i-th unit vector) to obtain $f_j(x)$ satisfying (7) with K_j instead of K. Let

$$P(x) = m(x_0) + \sum_{j \in \{1, \ldots, N\}^d} d_j \cdot f_j(x).$$

Then we can conclude from (7)

$$|P(x) - S(x)| \leq c_{17} \cdot N^{-p}$$

for all $x \in A$ which are not contained in

$$\cup_{i=1,\ldots,d} \cup_{j \in \{1,\ldots,N\}^d} \left\{ x \in \mathbb{R}^d \ : \ |x^{(i)} - x_j^{(i)}| < \eta/(6 \cdot d \cdot N) \right\}. \tag{9}$$

By shifting the positions of the x_j's in the i-th component we can construct $\lceil d/\eta \rceil$ disjoint versions of

$$\cup_{j \in \{1,\ldots,N\}^d} \left\{ x \in \mathbb{R}^d \ : \ |x^{(i)} - x_j^{(i)}| < \eta/(6 \cdot d \cdot N) \right\},$$

and since the sum of the ν–measures of these sets is less than or equal to one, at least one of them must have measure less than or equal to η/d. Consequently we can shift the x_j's such that (9) has ν–measure less than η. This together with (8) implies the assertion. □

ACKNOWLEDGMENT

The authors wish to thank two anonymous reviewers for their valuable comments and suggestions.

REFERENCES

[1] M. Anthony and P. L. Bartlett, *Neural Networks and Learning: Theoretical Foundations*. Cambridge, U.K.: Cambridge Univ. Press, 1999.
[2] C. de Boor, *A Practical Guide to Splines*. New York, NY, USA: Springer, 1978.
[3] A. M. Bagirov, C. Clausen, and M. Kohler, "Estimation of a regression function by maxima of minima of linear functions,"*IEEE Trans. Inf. Theory*, vol. 55, no. 2, pp. 833–845, Feb. 2009.
[4] A. R. Barron, "Complexity regularization with application to artificial neural networks," in *Nonparametric Functional Estimation and Related Topics*, G. Roussas, Ed. Dordrecht, The Netherlands: Kluwer Publishers, 1991.
[5] A. R. Barron, "Universal approximation bounds for superpositions of a sigmoidal function," *IEEE Trans. Inf. Theory*, vol. 39, no. 3, pp. 930–944, May 1993.
[6] L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*. New York, NY, USA: Springer, 1996.
[7] J. H. Friedman, "Multivariate adaptive regression splines," *Ann. Statist.*, vol. 19, no. 1, pp. 1–141, Mar. 1991.
[8] J. H. Friedman and W. Stuetzle, "Projection pursuit regression," *J. Amer. Statist. Assoc.*, vol. 76, no. 376, pp. 817–823, 1981.
[9] L. Györfi, M. Kohler, A. Krzyżak, and H. Walk, *A Distribution-Free Theory of Nonparametric Regression*. New York, NY, USA: Springer-Verlag, 2002.
[10] W. Härdle, P. Hall, and H. Ichimura, "Optimal smoothing in single-index models," *Ann. Statist.*, vol. 21, no. 1, pp. 157–178, 1993.
[11] W. Härdle, M. Müller, S. Sperlich, and A. Werwatz, *Nonparametric and Semiparametric Models*. New York, NY, USA: Springer-Verlag, 2004.

[12] W. Härdle and T. M. Stoker, "Investigating smooth multiple regression by the method of average derivatives," *J. Amer. Statist. Assoc.*, vol. 84, no. 408, pp. 986–995, 1989.
[13] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York, NY, USA: Springer, 2011.
[14] S. O. Haykin, *Neural Networks and Learning Machines*, 3rd ed. New York, NY, USA: Prentice-Hall, 2008.
[15] J. Hertz, A. Krogh, and R. G. Palmer, *Introduction to the Theory of Neural Computation*. Redwood City, CA, USA: Addison-Wesley, 1991.
[16] J. L. Horowitz and E. Mammen, "Rate-optimal estimation for a general class of nonparametric regression models with unknown link functions," *Ann. Statist.*, vol. 35, no. 6, pp. 2589–2619, 2007.
[17] M. Kohler, "Inequalities for uniform deviations of averages from expectations with applications to nonparametric regression," *J. Statist. Planning Inference*, vol. 89, nos. 1–2, pp. 1–23, 2000.
[18] M. Kohler and A. Krzyżak, "Adaptive regression estimation with multilayer feedforward neural networks," *J. Nonparam. Statist.*, vol. 17, no. 89, pp. 891–913, 2005.
[19] E. Kong and Y. Xia, "Variable selection for the single-index model," *Biometrika*, vol. 94, no. 1, pp. 217–229, 2007.
[20] G. Lugosi and K. K. Zeger, "Nonparametric estimation via empirical risk minimization," *IEEE Trans. Inf. Theory*, vol. 41, no. 3, pp. 677–687, May 1995.
[21] D. F. McCaffrey and A. R. Gallant, "Convergence rates for single hidden layer feedforward networks," *Neural Networks*, vol. 7, no. 1, pp. 147–158, 1994.
[22] H. N. Mhaskar, "Approximation properties of multilayer feedforward artificial neural network," *Adv. Comput. Math.*, vol. 1, no. 1, pp. 61–80, 1993.
[23] J. Mielniczuk and J. Tyrcha, "Consistency of multilayer perceptron regression estimators," *Neural Netw.*, vol. 6, no. 7, pp. 1019–1022, 1993.
[24] B. D. Ripley, *Pattern Recognition and Neural Networks*. Cambridge, U.K.: Cambridge Univ. Press, 2008.
[25] C. J. Stone, "Optimal global rates of convergence for nonparametric regression," *Ann. Statist.*, vol. 10, no. 4, pp. 1040–1053, 1982.
[26] C. J. Stone, "Additive regression and other nonparametric models," *Ann. Statist.*, vol. 13, no. 2, pp. 689–705, 1985.
[27] C. J. Stone, "The use of polynomial splines and their tensor products in multivariate function estimation," *Ann. Statist.*, vol. 22, no. 1, pp. 118–184, 1994.
[28] C. J. Stone, M. H. Hansen, C. Kooperberg, and Y. K. Truong, "Polynomial splines and their tensor product in extended linear modelling," *Ann. Statist.*, vol. 25, no. 4, pp. 1371–1470, 1997.
[29] Y. Yu and D. Ruppert, "Penalized spline estimation for partially linear single-index models," *J. Amer. Statist. Assoc.*, vol. 97, no. 460, pp. 1042–1054, Jan. 2002.

Michael Kohler received diploma degrees in computer science and mathematics from the University of Stuttgart in 1995, and a Ph.D. degree in mathematics from the University of Stuttgart in 1997. In 1998 he worked as a Visiting Scientist at the Stanford University, Stanford, USA. From 2005 till 2007 he was a Professor of Applied Mathematics at the University of Saarbrücken, Germany, since 2007 he is a Professor of Mathematical Statistics at the Technische Universität Darmstadt, Germany. He co-authored with L. Györfi, A. Krzyżak and H. Walk the book *A Distribution-Free Theory of Nonparametric Regression* (New York: Springer, 2002). His main research interests are in the area of nonparametric statistics, especially curve estimation and uncertainty quantification.

Adam Krzyżak (F'12) received the M.Sc. and Ph.D. degrees in computer engineering from the Technical University of Wroclaw, Poland, in 1977 and 1980, respectively, and D.Sc. degree (habilitation) in computer engineering from the Warsaw University of Technology, Poland in 1998. In 2003 he received the Title of Professor from the President of the Republic of Poland. Since 1983, he has been with the Department of Computer Science, Concordia University, Montreal, Canada, where he is currently a Professor. In 1983, he held International Scientific Exchange Award in the School of Computer Science, McGill University, Montreal, Canada, in 1991, Vineberg Memorial Fellowship at Technion-Israel Institute of Technology and, in 1992, Humboldt Research Fellowship at the University of Erlangen-Nürnberg, Germany. He visited the University of California Irvine, Information Systems Laboratory at Stanford University, Riken Frontiers Research Laboratory, Japan, Stuttgart University, Technical University of Berlin, University of Saarlandes and Technical University Darmstadt. He published nearly 300 papers on neural networks, pattern recognition, nonparametric estimation, image processing, computer vision and control. He has been an associate editor of IEEE TRANSACTIONS ON NEURAL NETWORKS and IEEE TRANSACTIONS ON INFORMATION THEORY and is presently a member of the editorial board of the *Pattern Recognition Journal*. He was co-editor of the book Computer Vision and Pattern Recognition (Singapore: World Scientic, 1989) and is co-author of the book *A Distribution-Free Theory of Nonparametric Regression*, Springer-Verlag, 2002. He has served among others on the program committees of Vision Interface Conference, International Conference on Document Processing and Applications, and International Conference on Computer Vision, Pattern Recognition and Image Processing, International Conference on Pattern Recognition and has been co-chair of Program Committee of the 10-th IEEE International Conference on Advanced Video and Signal-Based Surveillance. He co-organized a workshop at NIPS'94 Conference and was a session organizer at The World Congress of Nonlinear Analysts in 2000, 2004 and 2008. He is a Fellow of the IEEE.