Nonparametric Estimation of a Function From Noiseless Observations at Random Points

Benedikt Bauer[1,*], Luc Devroye[2], Michael Kohler[1,*], Adam Krzyżak[3] and Harro Walk[4]

[1] *Fachbereich Mathematik, Technische Universität Darmstadt, 64289 Darmstadt, Germany, email: bbauer@mathematik.tu-darmstadt.de, kohler@mathematik.tu-darmstadt.de*

[2] *School of Computer Science, McGill University, Montreal H3A 2K6, Canada, email: lucdevroye@gmail.com*

[3] *Department of Computer Science and Software Engineering, Concordia University, Montreal H3G 1M8, Canada, email: krzyzak@cs.concordia.ca*

[4] *Fachbereich Mathematik, Universität Stuttgart, 70569 Stuttgart, Germany, email: walk@mathematik.uni-stuttgart.de.*

August 2, 2016

Abstract

In this paper we study the problem of estimating a function from n noiseless observations of function values at randomly chosen points. These points are independent copies of a random variable which has a density with respect to the Lebesgue-Borel measure. This density is bounded away from zero on the unit cube and vanishes outside. The function to be estimated is assumed to be (p, C)-smooth, i.e., (roughly speaking) it is p-times continuously differentiable. Our main results are that the supremum norm error of a suitably defined spline estimate is bounded in probability by $(\log(n)/n)^{p/d}$ for arbitrary p and d and that this rate of convergence is optimal in minimax sense.

Keywords: nonparametric regression without noise, rate of convergence, spline estimate, supremum norm error

*The authors would like to thank the German Research Foundation (DFG) for funding this project within the Collaborative Research Centre 805.

1. INTRODUCTION

1.1 Multivariate Scattered Data Approximation

Approximation problems in which the input data is a set of deterministic distinct points are so-called scattered data approximation problems which have been extensively studied in the literature. In a typical setting we are given a set of deterministic points $(x_i, y_i) \in [0,1]^d \times \mathbb{R}^d$ $(i = 1, \ldots, n)$ and try to find a function m from a given function space, e.g., a Sobolev space, that fits the data closely. In scattered data approximation the points are not assumed to occupy a regular grid but rather are scattered around the space making the reconstruction problem difficult. The most popular approaches include the moving least squares approximation (Lancaster and Salkauskas (1981); Farwig (1986); Levin (1998); Wendland (2001, 2005); Joldes et al. (2015)), schemes based on radial basis functions or constant functions on spheres (Lazzaro and Montefusco (2002); Ohtake et al. (2005, 2006); Narcowich et al. (2006); Johnson et al. (2009)), multiquadric interpolants (Micchelli (1986)) and the smoothing spline approach. The latter one can be posed as the regularized least squares problem where one minimizes the criterion $\sum_{i=1}^{n}(m(x_i) - y_i)^2 + \lambda||m||_H^2$ over a class of functions H. The classes of functions include Beppo-Levi space (Johnson et al. (2009)) and Reproducing Kernel Hilbert Space (Gia et al. (2006)). In the moving least squares approach we seek function m^* which is a solution of the following minimization problem:

$$\min_{m \in P}\{\sum_{i=1}^{n}(m(x_i) - y_i)^2 w(x, x_i)\}, \tag{1}$$

where P is a finite-dimensional subspace (usually spanned by polynomials) of a space of continuous functions on a compact set Ω. Weight functions w are typically local, radial functions. It can be shown under mild conditions that the solution of problem (1) exists and is unique (Wendland (2001)). For the rate of approximation define the separation distance q_X and the mesh norm $h_{X,\Omega}$ as follows:

$$q_X = \frac{1}{2} \min_{1 \leq j < k \leq n} ||x_j - x_k|| \quad \text{and} \quad h_{X,\Omega} = \sup_{x \in \Omega} \min_{j \in \{1,\dots,n\}} \|x - x_j\|,$$

where $\|x\|$ denotes the Euclidean norm of $x \in \mathbb{R}^d$. Assume that a global constant c_1 exists such that the data separation condition

$$q_X \leq h_{X,\Omega} \leq c_1 \cdot q_X \tag{2}$$

holds on the data set. Then under the condition that Ω is compact and satisfies the so-called cone condition we get for $f \in C^p(\Omega)$ the approximation bound $\|m - m^*\|_{\infty,\Omega} \leq c_2 \cdot h^p_{X,\Omega}$, see, Wendland (2001, 2005). Hence if $x_1, \dots, x_n$ are scattered approximately evenly in $[0,1]^d$, we get

$$\|m - m^*\|_{\infty,[0,1]^d} \leq c_3 \cdot n^{-p/d}. \tag{3}$$

The approximation error bounds for the radial basis function interpolations may be found in Wendland (2005) and Madych and Nelson (1992).

1.2 The Problem Studied in this Paper

In practice it is not clear especially in high dimensions at which locations an estimated function should be sampled. A simple but effective way is to generate sampling points randomly from the uniform distribution on a ball or cube. The rest of the paper will be devoted to estimation of an unknown function m observed at such random scattered data. Our main question is how the error bound in (3) changes in this case. Obviously the result in (3) is not applicable in this case since condition (2) does not hold. Nevertheless it is natural to conjecture that a bound similar to (3) should hold for suitably defined estimates, even if the data points are randomly and not deterministically distributed. However, it is not clear how the definition of the estimates should be changed in order to be able to show such a result.

To formulate our problem precisely, let X, X_1, $\ldots$, X_n be independent and identically distributed random variables with values in $[0,1]^d$ and let $m : [0,1]^d \to \mathbb{R}$ be a (measurable) function. Given the data $\mathcal{D}_n = \{(X_1, m(X_1)), \ldots, (X_n, m(X_n))\}$ we are interested in constructing an estimate $m_n = m_n(\cdot, \mathcal{D}_n) : \mathbb{R}^d \to \mathbb{R}$ such that the supremum norm error $\|m_n - m\|_{\infty,[0,1]^d} = \sup_{x \in [0,1]^d} |m_n(x) - m(x)|$ is small.

1.3 Main Results

It is well-known that we need smoothness assumptions on m in order to derive nontrivial results on the rate of convergence of the global error of a function estimate (cf., e.g., Györfi et al. 2002, Theorem 3.1). In the sequel we assume that m is (p, C)–smooth for some $p = k + s$ for some $k \in \mathbb{N}_0$, $s \in (0, 1]$ and $C > 0$, i.e., (roughly speaking, see below for the exact definition) it is p-times continuously differentiable. Furthermore we will assume throughout this paper that there exists a constant $c_4 > 0$ such that

$$\mathbf{P}_X(S_r(x)) > c_4 \cdot r^d$$

does hold for all $x \in [0,1]^d$ and all $0 < r \leq 1$, where $S_r(x)$ denotes the (closed) ball of radius r around x. (This condition is in particular satisfied if X has a density with respect to the Lebesgue-Borel measure which is bounded away from zero on $[0,1]^d$.) We will show that in this case we can construct for an arbitrary $p > 0$ a spline estimate $m_n = m_n(\cdot, \mathcal{D}_n)$ such that

$$\|m_n - m\|_{\infty,[0,1]^d} = O_{\mathbf{P}}\left(\left(\frac{\log n}{n}\right)^{p/d}\right), \tag{4}$$

where we write $Z_n = O_{\mathbf{P}}(Y_n)$ if the nonnegative random variables Z_n and Y_n satisfy $\lim_{c\to\infty} \limsup_{n\to\infty} \mathbf{P}\{Z_n > c \cdot Y_n\} = 0$. Furthermore we show that the above rate of convergence is optimal in some minimax sense.

1.4 Discussion of Related Results

The estimation problem considered in this paper is a regression estimation problem without noise in the dependent variable. The case with noise in the dependent variable has been studied much more extensively in the literature. The common strategies comprise kernel regression estimates (cf., e.g., Nadaraya (1964, 1970), Watson (1964), Devroye and Wagner (1980), Stone (1977, 1982), Devroye and Krzyżak (1989)), partitioning regression estimates (cf., e.g., Györfi (1981), Beirlant and Györfi (1998)), nearest neighbor regression estimates (cf., e.g., Devroye (1982), Devroye et al. (1994)), least squares estimates (cf., e.g., Lugosi and Zeger (1995), Kohler (2000)) and smoothing spline estimates (cf., e.g., Wahba (1990), Kohler and Krzyżak (2001)).

Minimax rate of convergence results for the global errors of such estimates have been derived in Stone (1982). In particular it was shown there, that in case of the L_2 error and a (p, C)–smooth regression function the optimal rate of convergence is

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

In the setting of fixed design regression estimation it was analyzed in Kohler (2014) how the above rates of convergence changes if there is no noise in the dependent variable. The main result there are that for suitably defined spline estimates the supremum norm error converges to zero with the rate

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

(which follows already from the bound (3)) and that this rate of convergence is optimal in some minimax sense.

For the problem studied in this article Kohler and Krzyżak (2013) showed that the expected L_1-error of a nearest-neighbor estimate achieves the rate of convergence $n^{-\frac{p}{d}}$ in case $p \leq 1$. For $d = 1$ there was also an estimate constructed which achieves this rate of convergence for arbitrary p.

In contrast, our results consider the supremum norm error and are applicable for general p and d. Here it is natural to conjecture that results like (3) lead to bounds like (4), however it is not clear how one can construct an estimate for random scattered data achieving the rate (4).

1.5 Notation

The sets of natural numbers, natural numbers including 0, and real numbers are denoted by $\mathbb{N}$, $\mathbb{N}_0$ and $\mathbb{R}$, resp. The Euclidean norm of $x \in \mathbb{R}^d$ is denoted by $\|x\|$. For $f : \mathbb{R}^d \to \mathbb{R}$ the expression $\|f\|_\infty = \sup_{x \in \mathbb{R}^d} |f(x)|$ is its supremum norm, and the supremum norm of f on a set $A \subseteq \mathbb{R}^d$ is denoted by $\|f\|_{\infty,A} = \sup_{x \in A} |f(x)|$. $S_r(x)$ is the (closed) ball of radius r around x. A function $f : \mathbb{R}^d \to \mathbb{R}$ is called (p, C)-smooth, where $C > 0$ and $p = k + s$ with $k \in \mathbb{N}_0$ and $s \in (0, 1]$ hold, if for every $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}_0^d$ with $\sum_{j=1}^d \alpha_j = k$ the partial derivative $\frac{\partial^k f}{\partial x_1^{\alpha_1} \ldots \partial x_d^{\alpha_d}}$ exists and satisfies

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

for all $x, z \in \mathbb{R}^d$. For $z \in \mathbb{R}$ we denote the smallest integer greater than or equal to z by by $\lceil z \rceil$, and $\lfloor z \rfloor$ denotes the largest integer less than or equal to z.
If not otherwise stated, any c_i with $i \in \mathbb{N}$ here and in the following symbolizes a real constant, which is nonnegative.

1.6 Outline

In Section 2 we define our estimates, the main results are presented in Section 3, several simulations are presented in Section 4, and Section 5 contains the proofs. An elementary bound on a probability needed in one of our proofs is given in the appendix.

2. DEFINITION OF THE ESTIMATES

2.1 The Main Idea

In the sequel we want to estimate a function from noiseless observations of function values at randomly scattered points. Here we want to fit a function from a given class of functions to our data. For this we could use, e.g., the principle of the (penalized) least squares, however this would not take advantage of the fact that our observations are noiseless and hence highly trustworthy. Otherwise we could try to interpolate our function values, however this will cause problems since our points will be irregularly spaced and consequently some areas of our sample space will need much more degrees of freedom of our interpolant than other areas.

The key idea introduced in this paper is to find an estimate such that the maximal distance between its values and the observed function values is smaller than some threshold. More precisely, well will choose $\delta_n > 0$ and a suitable function space $\mathcal{F}_n$, and will choose an estimate m_n such that $m_n \in \mathcal{F}_n$ and $|m_n(X_i) - m(X_i)| \leq \delta_n$ for all $i \in \{1, \ldots, n\}$. In case that $\mathcal{F}_n$ is a finite dimensional linear vector space of functions, linear programming can be used to compute such an estimate.

2.2 The Spline Estimate

In this subsection we assume that $m : \mathbb{R}^d \to \mathbb{R}$ is (p, C)–smooth for $C > 0$ and $p = k + s$ with $k \in \mathbb{N}_0$ and $s \in (0, 1]$. In order to define the spline estimate, a B-spline basis of functions with compact support, which spans the space of polynomial splines (i.e., of piecewise polynomials satisfying a global smoothness condition) on $[0, 1]^d$, is introduced.

Definition 1. *Choose $K \in \mathbb{N}$, $M \in \mathbb{N}_0$ and set $u_j = j/K$ ($j \in \{-M, \dots, K + M\}$). The (univariate)* B-splines *$B_{j,l} : \mathbb{R} \to \mathbb{R}$ of degree l are recursively defined by*

(i)

$$B_{j,0}(x) = \begin{cases} 1, & x \in [u_j, u_{j+1}) \\ 0, & x \notin [u_j, u_{j+1}) \end{cases}$$

for $j = -M, \dots, K + M - 1$ and

(ii)

$$B_{j,l+1}(x) = \frac{x - u_j}{u_{j+l+1} - u_j} B_{j,l}(x) + \frac{u_{j+l+2} - x}{u_{j+l+2} - u_{j+1}} B_{j+1,l}(x)$$

for $j = -M, \dots, K + M - l - 2$ and $l = 0, \dots, M - 1$.

The sequence $(u_j)_{j=-M,\dots,K+M}$ is called knot sequence *and M is called* degree *of the B-splines.*

In order to be able to define spaces of multivariate funtions, the univariate B-splines are combined to form multivariate tensor product B-splines.

Definition 2. *Choose $K \in \mathbb{N}$ and $M \in \mathbb{N}_0$. For $\mathbf{j} = (j_1, \dots, j_d) \in \{-M, \dots, K + M\}^d$ the* tensor product B-spline *$B_{\mathbf{j},M} : \mathbb{R}^d \to \mathbb{R}$ is defined by*

$$B_{\mathbf{j},M}(x) = B_{j_1,M}\left(x^{(1)}\right) \cdot \ldots \cdot B_{j_d,M}\left(x^{(d)}\right).$$

For $M \in \mathbb{N}$ and $K = K_n = \left\lfloor c_5 \cdot (n/\log n)^{1/d} \right\rfloor$ with a certain $c_5 > 0$ let $\{B_{\mathbf{j},M} : \mathbf{j} \in \{-M, \ldots, K_n - 1\}^d\}$ be the corresponding tensor product B-splines. We define our estimate by

$$m_n(x) = \sum_{\mathbf{j} \in \{-M,\ldots,K_n-1\}^d} \hat{c}_{\mathbf{j}} \cdot B_{\mathbf{j},M}(x) \quad \left(x \in [0,1]^d\right) \tag{5}$$

with coefficients $\hat{c}_{\mathbf{j}} \in \mathbb{R}$ such that m_n approximates the observed data. If the spline degree $M \in \mathbb{N}$ fulfills the condition $M \geq k$, then it follows from Theorem 1 in Kohler (2014) that it is possible to choose these coefficients such that $|m_n(x) - m(x)| \leq c_6 \cdot K_n^{-p}$ holds for a constant $c_6 > 0$ depending only on d, M, p and C. So we set $\delta_n = c_7 \cdot K_n^{-p}$ for a suitably chosen $c_7 > 0$ and choose the coefficients $\hat{c}_{\mathbf{j}}$ such that the following n inequalities are satisfied:

$$|m_n(X_i) - m(X_i)| \leq \delta_n \quad (i = 1, \ldots, n)\,. \tag{6}$$

For n sufficiently large and m (p, C)–smooth, the solution space of this system of inequalities must be non-empty because of the above-mentioned result in Kohler (2014). Linear programming can be used to compute the coefficients $\hat{c}_{\mathbf{j}} \in \mathbb{R}$.

3. MAIN RESULTS

We start with deriving an upper bound on the rate of convergence of our estimate (5).

Theorem 1. *Let X, X_1, X_2 be indepenent and identically distributed random variables with values in $\mathbb{R}^d$. Assume that there exists a constant $c_8 > 0$ such that*

$$\mathbf{P}_X(S_r(x)) > c_8 \cdot r^d$$

does hold for all $x \in [0,1]^d$ and all $0 < r \leq 1$. Let $m : \mathbb{R}^d \to \mathbb{R}$ be (p, C)–smooth for $C > 0$

and $p = k + s$ *with* $k \in \mathbb{N}_0$ *and* $s \in (0,1]$. *Choose* $M \in \mathbb{N}$ *with* $M \geq k$ *and set*

$$K_n = \left\lceil c_9 \cdot \left(\frac{n}{\log n} \right)^{\frac{1}{d}} \right\rceil \quad and \quad \delta_n = c_{10} \cdot \left(\frac{\log n}{n} \right)^{p/d}.$$

Let m_n *be defined by (5) and (6). Then for* $c_9 > 0$ *sufficiently small and* $c_{10} > 0$ *sufficiently large we have*

$$\|m_n - m\|_{\infty,[0,1]^d} = O_{\mathbf{P}} \left(\left(\frac{\log n}{n} \right)^{p/d} \right).$$

Remark 1. The proof of Theorem 1 implies that the bound on the probability in Theorem 1 holds uniformly over the class of (p, C)–smooth functions (for a fixed distribution of X satisfying the assumptions in Theorem 1, e.g., for uniform distribution on the unit cube). More precisely, we can conclude from the proof of Theorem 1, that our estimate satisfies for some $c_{11} > 0$

$$\limsup_{n \to \infty} \sup_{m \in \mathcal{F}^{(p,C)}} \mathbf{P} \left\{ \|m_n - m\|_{\infty,[0,1]^d} \geq c_{11} \cdot \left(\frac{\log n}{n} \right)^{p/d} \right\} = 0,$$

where $\mathcal{F}^{(p,C)}$ denotes the set of all (p, C)-smooth functions $m : \mathbb{R}^d \to \mathbb{R}$.

Next we show that the rate of convergence in Theorem 1 as formulated in Remark 1 is optimal whenever estimating (p, C)–smooth functions from noiseless observations at random points.

Theorem 2. *Let* $p = k + s$ *for some* $k \in \mathbb{N}_0$ *and* $s \in (0,1]$ *and let* $C > 0$. *Let* $\mathcal{F}^{(p,C)}$ *denote the set of all* (p, C)*-smooth functions* $m : \mathbb{R}^d \to \mathbb{R}$ *and let* $X_1, \dots, X_n$ *be independent and uniformly distributed on* $[0,1]^d$. *Then there is a constant* $c_{12} > 0$ *such that*

$$\liminf_{n \to \infty} \inf_{m_n} \sup_{m \in \mathcal{F}^{(p,C)}} \mathbf{P} \left\{ \|m_n - m\|_{\infty,[0,1]^d} \geq c_{12} \cdot \left(\frac{\log n}{n} \right)^{\frac{p}{d}} \right\} > 0$$

holds.

4. APPLICATION TO SIMULATED DATA

In this section we apply the estimate developed in the previous section to simulated data and compare the results with conventional estimates using the statistics package *R*.
For this purpose, we consider three competitive approaches. The first one is interpolation with radial basis functions presented in Lazarro and Montefusco (2002), where authors use Wendland's compactly supported radial basis function $\phi(r) = (1-r)_+^6 \cdot (35r^2 + 18r + 3)$. The second approach to which we compare our estimate is the moving least squares estimate with the second order polynomial basis and a quartic weight function as described in Joldes et al. (2015), where we scale the radius of influence they used with respect to the size of our estimation area and the sample size. Instead of their modification we use the Moore-Penrose generalized inverse of the matrix if it is singular because this yields to better results and works even for very ill-conditioned matrices. The third approach is thin plate spline estimate whose smoothing parameter is chosen by the generalized cross validation as implemented by the routine *Tps()* of the library *fields* in *R*.
The parameters M and K of our spline estimate defined in (6) are chosen adaptively by cross validation allowing values from 1 to M_{max} and K_{max}, respectively. M_{max} and K_{max} can take values up to 5 and 25, respectively depending on the examples, although the set of possible choices was reduced for some settings if several test runs showed that the whole range of choices is not needed. The parameter δ_n is chosen as the smallest possible value in $\{2^i/n : i = -50, \dots, 30\}$ such that a solution of the linear program exists.

Table 1 shows the results arising from our experiments. Random variable X is uniformly distributed on $[0,1]^2$ and we try six different test functions $m_i : [0,1]^2 \to \mathbb{R}$ $(i = 1, \dots, 6)$

Table 1: Median (IQR) of the errors of the estimates for $m_1, m_2, m_3, m_4, m_5, m_6$

function $\mathbf{m_1}$	$n = 50$	$n = 100$	$n = 200$
spline estimate	**8.3E-10 (8.5E-10)**	**1.5E-10 (1.7E-10)**	**2.5E-11 (1.0E-11)**
RBF interpolant	5.4E-1 (3.2E-1)	3.0E-1 (7.3E-2)	1.5E-1 (1.5E-1)
MLS estimate	5.6E-2 (3.5E-2)	4.0E-2 (1.5E-2)	3.1E-2 (5.0E-3)
thin plate spline	3.5E-1 (1.6E-1)	2.0E-1 (5.7E-2)	1.3E-1 (9.4E-2)
function $\mathbf{m_2}$	$n = 50$	$n = 100$	$n = 200$
spline estimate	**1.8E-1 (3.4E-1)**	1.1E-1 (1.1E-1)	**5.0E-2 (7.0E-2)**
RBF interpolant	1.2E0 (7.4E-1)	5.7E-1 (4.0E-1)	1.8E-1 (1.5E-1)
MLS estimate	2.2E-1 (1.4E-1)	**9.4E-2 (4.2E-2)**	9.0E-2 (2.0E-2)
thin plate spline	2.2E-1 (9.8E-2)	1.4E-1 (7.3E-2)	7.5E-2 (1.9E-2)
function $\mathbf{m_3}$	$n = 50$	$n = 100$	$n = 200$
spline estimate	1.7E-1 (1.8E-1)	**4.4E-2 (6.4E-2)**	**2.2E-2 (2.1E-2)**
RBF interpolant	3.8E-1 (2.8E-1)	1.6E-1 (2.2E-1)	8.4E-2 (6.2E-2)
MLS estimate	**8.9E-2 (3.0E-2)**	5.3E-2 (1.5E-2)	4.7E-2 (1.2E-2)
thin plate spline	1.4E-1 (8.8E-2)	8.4E-2 (3.2E-2)	5.8E-2 (2.3E-2)
function $\mathbf{m_4}$	$n = 50$	$n = 100$	$n = 200$
spline estimate	1.4E-1 (1.7E-1)	**8.2E-3 (8.2E-3)**	**4.3E-3 (3.9E-3)**
RBF interpolant	**5.5E-2 (6.6E-2)**	1.8E-2 (3.3E-2)	5.2E-3 (3.9E-3)
MLS estimate	8.3E-2 (2.7E-2)	3.2E-2 (7.1E-3)	2.7E-2 (6.5E-3)
thin plate spline	1.0E-1 (7.4E-2)	5.5E-2 (5.5E-2)	2.2E-2 (7.5E-3)
function $\mathbf{m_5}$	$n = 50$	$n = 100$	$n = 200$
spline estimate	**2.8E-1 (3.0E-1)**	2.0E-1 (1.8E-1)	**5.4E-2 (7.6E-2)**
RBF interpolant	3.2E-1 (1.3E-1)	**1.6E-1 (1.1E-1)**	6.1E-2 (2.9E-2)
MLS estimate	7.0E-1 (3.7E-1)	3.0E-1 (6.2E-2)	2.4E-1 (1.1E-1)
thin plate spline	7.8E-1 (3.3E-1)	4.5E-1 (2.6E-1)	1.9E-1 (8.8E-2)
function $\mathbf{m_6}$	$n = 50$	$n = 100$	$n = 200$
spline estimate	3.1E-1 (2.2E-1)	2.6E-1 (1.3E-1)	1.8E-1 (1.1E-1)
RBF interpolant	2.4E-1 (1.2E-1)	1.4E-1 (9.1E-2)	5.6E-2 (1.7E-2)
MLS estimate	1.3E-1 (4.8E-2)	8.8E-2 (7.0E-3)	8.5E-2 (7.2E-3)
thin plate spline	**1.0E-1 (2.7E-2)**	**7.2E-2 (2.8E-2)**	**4.0E-2 (1.6E-2)**

with different degrees of (p, C)–smoothness as illustrated in Figure 1 and defined as follows.

$$\begin{aligned}
m_1(x_1, x_2) &= 3 \cdot x_1^2 \cdot x_2 - x_2^3, \\
m_2(x_1, x_2) &= 2 \cdot \exp(-5 \cdot (x_1 - 0.7)^2) - \exp(-5 \cdot (x_1 - 0.4)^2) - 3 \cdot x_2 + 5, \\
m_3(x_1, x_2) &= \frac{1}{x_1 + x_2^3 + 0.5} \\
m_4(x_1, x_2) &= \exp(-3 \cdot ((x_1 - 0.75)^2 + (x_2 - 0.75)^2)), \\
m_5(x_1, x_2) &= \sin(2 \cdot \pi \cdot x_1) \cdot \cos(\pi \cdot x_2), \\
m_6(x_1, x_2) &= \min\{1 - x_2, 2 \cdot x_1 - 0.5\}.
\end{aligned}$$

The estimates under consideration are computed for different numbers ($n = 50, 100, 200$) of independent realizations of X and their corresponding function values. Since the results of the simulations depend on the randomly chosen data points, we compute the estimates repeatedly ($N = 50$ times) for regenerated realizations of X and examine the median (plus interquartile range IQR) of the supremum errors with respect to an equidistant grid of width 0.02 on $[0, 1]^2$. Examination of the results shows that, on the one hand, our general spline estimate clearly outperforms the comparative estimates in the polynomial case of m_1 (even for small sample sizes), which could have been expected because our estimate consists of piecewise polynomials. In addition to that, it has the smallest median error for increasing sample sizes in the moderately smooth cases of m_2 to m_5 (with only a comparatively small advantage in the volatile case of m_5). On the other hand, it is relatively bad for all of the considered sample sizes in the edged case of m_6, which is not even differentiable, but it improves steadily. All of these observations go well with the convergence rates deduced in the previous sections, which depend on the (p, C)–smoothness of the function (cf. Theorem 1).

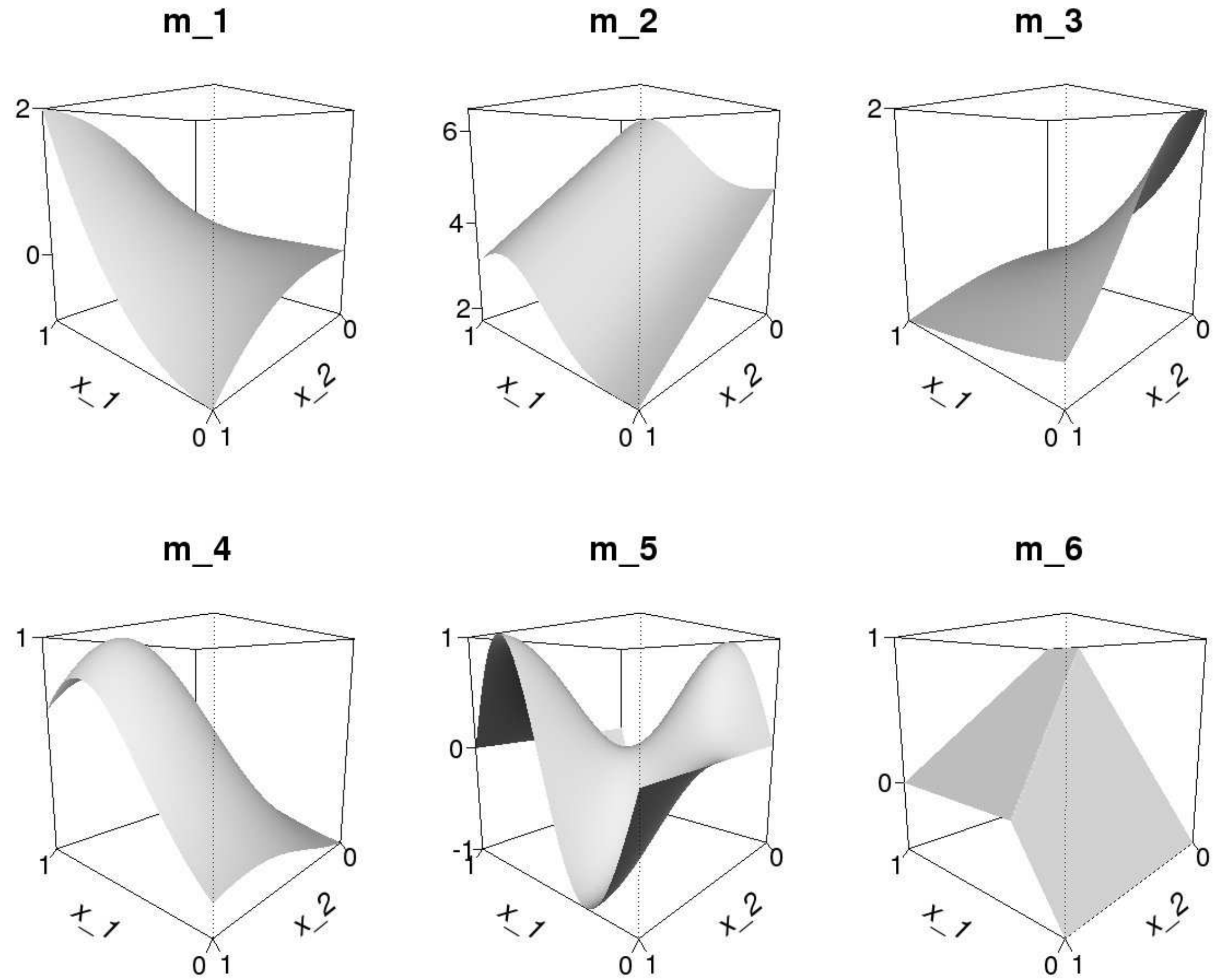

Figure 1: Behavior of the test functions $m_1, m_2, m_3, m_4, m_5, m_6$.

5. PROOFS

5.1 Proof of Theorem 1

For the proof of Theorem 1 we need the following lemmata.

Lemma 1. *Let* Π *be the ring of all polynomials* $p : \mathbb{R}^d \to \mathbb{R}$ *in* d *variables and let* $\mathcal{P}$ *be a finite–dimensional subspace of* Π *with dimension* $\dim \mathcal{P}$*. For almost any set* $X \subset \mathbb{R}^d$ *of interpolations sites with* $|X| = \dim \mathcal{P}$ *there is an unique* $p \in \mathcal{P}$ *which fulfills* $p(x) = y_x$ *for all* $x \in X$ *and arbitrarily chosen values* $y_x \in \mathbb{R}$*.*

Proof. The assertion of the lemma follows immediately from Proposition 4 in Jetter et al. (2006). $\square$

Lemma 2. *Assume that the distribution of the i.i.d. random variables* $X, X_1, \ldots, X_n$ *satisfies* $\mathbf{P}_X(S_\varepsilon(x)) \geq c_{13} \cdot \varepsilon^d$ *for all* $x \in [0,1]^d$ *and* $\varepsilon \in (0,1]$ *where* $c_{13} > 0$ *is a constant and* $S_\varepsilon(x)$ *is the closed ball around* x *with radius* ε*. Let*

$$K_n = \left\lceil c_9 \cdot \left(\frac{n}{\log n} \right)^{\frac{1}{d}} \right\rceil .$$

Let $B_1, \ldots, B_{K_n^d}$ *be* K_n^d *balls with radius* $c_{14} \cdot \frac{1}{K_n}$*, whose centers lie in* $[0,1]^d$*. For any* $r > 0$*, there is a sufficiently small* $c_9 := c_9(r, c_{13}, c_{14}) > 0$*, such that*

$$\mathbf{P}\left\{ \forall j \in \{1, \ldots, K_n^d\} \ \exists i \in \left\{1, \ldots, \left\lfloor \frac{n}{r} \right\rfloor \right\} : X_i \in B_j \right\} \to 1 \quad (n \to \infty).$$

Proof. We consider the complementary event of the above expression. By the union bound, the independence of $X_1, \ldots, X_{\lfloor \frac{n}{r} \rfloor}$ and the assumption on the distribution of X we get for sufficiently large n

$$\begin{aligned}
& \mathbf{P}\left\{ \exists j \in \{1, \ldots, K_n^d\} : X_1, \ldots, X_{\lfloor \frac{n}{r} \rfloor} \notin B_j \right\} \\
\leq & \sum_{j \in \{1, \ldots, K_n^d\}} (1 - \mathbf{P}_X(B_j))^{\lfloor \frac{n}{r} \rfloor} \\
\leq & \ K_n^d \cdot \max_{j \in \{1, \ldots, K_n^d\}} (1 - \mathbf{P}_X(B_j))^{\lfloor \frac{n}{r} \rfloor}
\end{aligned}$$

$$
\begin{aligned}
&\leq K_n^d \cdot \left(1 - c_{13} \cdot \frac{c_{14}^d}{K_n^d}\right)^{\lfloor \frac{n}{r} \rfloor} \\
&\leq K_n^d \cdot \exp\left(-c_{13} \cdot c_{14}^d \cdot \frac{n}{2 \cdot r \cdot K_n^d}\right) \\
&\leq 2^d \cdot c_9^d \cdot \frac{n}{\log n} \cdot \exp\left(-c_{13} \cdot c_{14}^d \cdot \frac{\log n}{2 \cdot r \cdot c_9^d}\right)
\end{aligned}
$$

For sufficiently small c_9 the right-hand side of the inequality above tends to zero for n tending to infinity. □

Lemma 3. *Let the random variables $X, X_1, \ldots, X_n$ and the parameter K_n be chosen as in Lemma 2. Let $r \in \mathbb{N}$ be an arbitrary constant and let $B_1, \ldots, B_{r \cdot K_n^d}$ be $r \cdot K_n^d$ balls with radius $c_{14} \cdot \frac{1}{K_n}$, whose centers lie in $[0,1]^d$. Then for $c_9 := c_9(r, c_{13}, c_{14}) > 0$ sufficiently small*

$$
\mathbf{P}\left\{\forall j \in \left\{1, \ldots, r \cdot K_n^d\right\} \ \exists i \in \{1, \ldots, n\} : X_i \in B_j\right\} \to 1 \quad (n \to \infty).
$$

Proof. At first, we note that

$$
\begin{aligned}
&\mathbf{P}\left\{\forall j \in \left\{1, \ldots, r \cdot K_n^d\right\} \ \exists i \in \{1, \ldots, n\} : X_i \in B_j\right\} \\
&\geq \prod_{k=1}^{r} \mathbf{P}\Bigg\{\forall j \in \left\{(k-1) \cdot K_n^d + 1, \ldots, k \cdot K_n^d\right\} \\
&\qquad\qquad \exists i \in \left\{(k-1) \cdot \left\lfloor \frac{n}{r} \right\rfloor + 1, \ldots, k \cdot \left\lfloor \frac{n}{r} \right\rfloor\right\} : X_i \in B_j\Bigg\}
\end{aligned}
$$

The assertion follows from the application of Lemma 2 for the inner expression. □

Proof of Theorem 1. Let $Q_{\mathbf{j}}(m) \in \mathbb{R}$ be the coefficients of the spline approximant of m in Theorem 1 in Kohler (2014) which ensures that

$$\bar{m}_n(x) = \sum_{\mathbf{j} \in \{-M,\dots,K_n-1\}^d} Q_{\mathbf{j}}(m) \cdot B_{\mathbf{j},M}(x)$$

fulfills

$$|\bar{m}_n(x) - m(x)| \leq c_{15} \cdot K_n^{-p} \tag{7}$$

for all $x \in [0,1]^d$ and a constant $c_{15} > 0$. This together with (6) implies

$$\begin{aligned} |m_n(X_i) - \bar{m}_n(X_i)| &\leq |m_n(X_i) - m(X_i)| + |\bar{m}_n(X_i) - m(X_i)| \\ &\leq \delta_n + c_{15} \cdot K_n^{-p} \\ &\leq c_{16} \cdot (\log n/n)^{p/d} \qquad (8) \end{aligned}$$

for n sufficiently large and an adequately chosen $c_{16} > 0$. Set

$$z_{\mathbf{j}} = \hat{c}_{\mathbf{j}} - Q_{\mathbf{j}}(m)$$

for every $\mathbf{j} \in \{-M,\dots,K_n-1\}^d$. Then

$$m_n(x) - \bar{m}_n(x) = \sum_{\mathbf{j} \in \{-M,\dots,K_n-1\}^d} z_{\mathbf{j}} \cdot B_{\mathbf{j},M}(x)$$

holds, so we can conclude

$$|m_n(x) - \bar{m}_n(x)| = \left| \sum_{\mathbf{j} \in \{-M,\dots,K_n-1\}^d} z_{\mathbf{j}} \cdot B_{\mathbf{j},M}(x) \right| \leq \max_{\mathbf{j} \in \{-M,\dots,K_n-1\}^d} |z_{\mathbf{j}}|$$

for all $x \in [0,1]^d$, because the B-splines are non-negative and sum up to 1 (cf., e.g., Lemma 15.2 in Györfi et al. (2002)). Combining this with the previous bounds we get

$$|m_n(x) - m(x)| \leq |m_n(x) - \bar{m}_n(x)| + |\bar{m}_n(x) - m(x)|$$

$$\leq \max_{\mathbf{j} \in \{-M,\ldots,K_n-1\}^d} |z_{\mathbf{j}}| + c_{17} \cdot (\log n/n)^{p/d}$$

and from now on it suffices to show that we have outside of an event, whose probability tends to zero for $n \to \infty$, $\max_{\mathbf{j} \in \{-M,\ldots,K_n-1\}^d} |z_{\mathbf{j}}| \leq c_{18} \cdot (\log n/n)^{p/d}$ for a certain $c_{18} > 0$.
By (8) our estimate fulfills

$$\sum_{\mathbf{j} \in \{-M,\ldots,K_n-1\}^d} z_{\mathbf{j}} \cdot B_{\mathbf{j},M}(X_i) = \varepsilon(i) \quad (i = 1, \ldots, n) \tag{9}$$

for an adequately chosen

$$\varepsilon(i) \in \left[-c_{16} \cdot \left(\frac{\log n}{n} \right)^{p/d}, c_{16} \cdot \left(\frac{\log n}{n} \right)^{p/d} \right] \quad \text{for all } i \in \{1, \ldots, n\}. \tag{10}$$

Next, consider a fixed d–dimensional spline node interval $A_{\mathbf{j}} = (u_{j_1}, u_{j_1+1}) \times \cdots \times (u_{j_d}, u_{j_d+1})$ for an arbitrarily chosen $\mathbf{j} \in \{-M, \ldots, K_n - 1\}^d$. Let $\mathcal{S}_{\mathbf{j}} \subseteq \{-M, \ldots, K_n - 1\}^d$ contain exactly those indices $\mathbf{k} = (k_1, \ldots, k_d)$ that fulfill

$$j_i - M \leq k_i \leq j_i \qquad (i \in \{1, \ldots, d\}).$$

If we put $(M+1)^d$ different values $x_1, \ldots, x_{(M+1)^d} \in A_{\mathbf{j}}$ in equations of the type (9), this leads to the linear system of equations

$$\sum_{\mathbf{k} \in \mathcal{S}_{\mathbf{j}}} z_{\mathbf{k}} \cdot B_{\mathbf{k},M}(x_i) = \varepsilon(i) \quad (i = 1, \ldots, (M+1)^d), \tag{11}$$

because the rest of the B–spline terms vanishes on $A_{\mathbf{j}}$. We will abbreviate (11) in matrix notation by $\mathbf{B}_{\mathbf{j}} \cdot \mathbf{z}_{\mathbf{j}} = \varepsilon_{\mathbf{j}}$. Because the remaining B–splines are polynomials on this set, (11) equals a polynomial interpolation problem on $A_{\mathbf{j}}$.
Since the B–splines are scaled regarding K_n and $A_{\mathbf{j}}$, we can consider this polynomial interpolation problem on $(0,1)^d$ instead of $A_{\mathbf{j}}$ and with B–splines respecting this larger node

distance. Due to the local linear independence of the B–splines (cf. Lemma 14.5 in Györfi et al. (2002)) the polynomials form a $(M+1)^d$–dimensional vector space. So according to Lemma 1 there is a set of distinct points $\tilde{x}_1, \ldots, \tilde{x}_{(M+1)^d} \in (0,1)^d$ (almost every set would work), such that this interpolation problem is uniquely solvable, which means $|\det(\mathbf{B_j})|$ is greater than zero. Moreover, the absolute value of the determinant of $\mathbf{B_j}$ is a continous function regarding the inputs $\tilde{x}_1, \ldots, \tilde{x}_{(M+1)^d}$ (since the B–splines are continuous functions of their arguments for degree greater than zero). So there is a closed ball with radius c_{19} around all of these values, where $|\det(\mathbf{B_j})| \geq c_{\min} > 0$ holds. Note that this argumentation is independent of the size of $A_\mathbf{j}$ (which depends on n), because the B–splines are scaled according to $A_\mathbf{j}$ by definition and the closed balls exist in a scaled version with radius $c_{19} \cdot \frac{1}{K_n^d}$ in $A_\mathbf{j}$. So $c_{\min}$ does not depend on n.
Due to Lemma 3 at least $(M+1)^d$ of the realizations X_i fall into the above-mentioned compact balls in $A_\mathbf{j}$ for sufficiently large n. We call these realizations $\tilde{X}_1, \ldots, \tilde{X}_{(M+1)^d}$. Their corresponding equations in (9) form a system like (11) which can be solved by Cramer's rule in the form of

$$z_\mathbf{k} = \frac{\det(\mathbf{B_j}(\mathbf{k}, \varepsilon_\mathbf{j}))}{\det(\mathbf{B_j})}$$

for all $\mathbf{k} \in \mathcal{S}_\mathbf{j}$, where $\mathbf{B_j}(\mathbf{k}, \varepsilon_\mathbf{j})$ symbolizes a version of $\mathbf{B_j}$, in which the column that belongs to $\mathbf{k}$ is replaced by $\varepsilon_\mathbf{j}$. The fact that the B–spline values and the determinant of $\mathbf{B_j}$ are bounded allows the conclusion

$$|z_\mathbf{k}| = \frac{|\det(\mathbf{B_j}(\mathbf{k}, \varepsilon_\mathbf{j}))|}{|\det(\mathbf{B_j})|} \leq \frac{c_{19}}{c_{\min}} \cdot \max_{i=1,\ldots,(M+1)^d} |\varepsilon(i)| \leq c_{18} \cdot \left(\frac{\log n}{n}\right)^{p/d} \tag{12}$$

because of (10). Since the above argumentation works for all of the $A_\mathbf{j}$ simultaneously (cf., Lemma 3), every $z_\mathbf{j}$ in (9) can be bounded by (12), and this implies the assertion. □

5.2 Proof of Theorem 2

Set $M_n = \left\lfloor (2 \cdot n/\log n)^{1/d} \right\rfloor$ and let $\{A_{n,j}\}_{j=1,\ldots,M_n^d}$ be a partition of $[0,1]^d$ into cubes of side length $\frac{1}{M_n}$. Choose a $(p, 2^{s-1}C)$-smooth function $g : \mathbb{R}^d \to \mathbb{R}$ (where s comes from the definition of the (p,C)–smoothness in the theorem) satisfying $supp(g) \subseteq \left(-\frac{1}{2}, \frac{1}{2}\right)^d$ and reaching a certain constant $c_{20} > 0$ on its support, i.e., satisfying $c_{20} = \sup_{x \in \mathbb{R}^d} g(x) = g(x_0) > 0$ for some $x_0 \in \left(-\frac{1}{2}, \frac{1}{2}\right)^d$. For $j \in \left\{1, \ldots, M_n^d\right\}$ let $a_{n,j}$ be the center of $A_{n,j}$ and set $g_{n,j}(x) = M_n^{-p} \cdot g\left(M_n \cdot (x - a_{n,j})\right)$. We define $m^{(c_n)} : \mathbb{R}^d \to \mathbb{R}$ by $m^{(c_n)} = \sum_{j=1}^{M_n^d} c_{n,j} \cdot g_{n,j}(x)$, where $c_n = (c_{n,j})_{j=1,\ldots,M_n^d} \in \{-1,1\}^{M_n^d}$.

The functions $m^{(c_n)}$ are (p,C)-smooth for all $(c_n) \in \{-1,1\}^{M_n^d}$ (cf., e.g., Györfi et al. (2002), proof of Theorem 3.2), hence we have

$$\left\{ m^{(c_n)} \,:\, c_n \in \{-1,1\}^{M_n^d} \right\} \subseteq \mathcal{F}^{(p,C)}. \tag{13}$$

Randomizing the coefficients of this type of functions we introduce random variables $C_{n,1}, \ldots, C_{n,M_n^d}$ which are independent from each other and from $X_1, \ldots, X_n$, such that $\mathbf{P}\{C_{n,k} = -1\} = \mathbf{P}\{C_{n,k} = 1\} = \frac{1}{2}$ for all $k = 1, \ldots, M_n^d$, and we set $C_n = \left(C_{n,1}, \ldots, C_{n,M_n^d}\right)$.
Using the relation $M_n \le (2 \cdot n/\log n)^{1/d}$, (13) allows the following bounding for an arbitrary estimate m_n:

$$
\begin{aligned}
&\sup_{m \in \mathcal{F}^{(p,C)}} \mathbf{P}\left\{ \|m_n(\cdot, (X_1, m(X_1)), \ldots, (X_n, m(X_n))) - m\|_{\infty,[0,1]^d} \ge c_{20} \cdot \left(\frac{\log n}{2 \cdot n}\right)^{\frac{p}{d}} \right\} \\
&\ge \sup_{c_n \in \{-1,1\}^{M_n^d}} \mathbf{P}\Bigg\{ \|m_n(\cdot, (X_1, m^{(c_n)}(X_1)), \ldots, (X_n, m^{(c_n)}(X_n))) - m^{(c_n)}\|_{\infty,[0,1]^d} \\
&\qquad\qquad \ge c_{20} \cdot M_n^{-p} \Bigg\} \\
&\ge \mathbf{P}\left\{ \|m_n(\cdot, (X_1, m^{(C_n)}(X_1)), \ldots, (X_n, m^{(C_n)}(X_n))) - m^{(C_n)}\|_{\infty,[0,1]^d} \ge c_{20} \cdot M_n^{-p} \right\} \\
&\ge \mathbf{P}\Bigg\{ \exists j \in \left\{1, \ldots, M_n^d\right\} : X_1 \notin A_{n,j}, \ldots, X_n \notin A_{n,j} \text{ and}
\end{aligned}
$$

$$|m_n(x_{0,j},(X_1,m^{(C_n)}(X_1)),\dots,(X_n,m^{(C_n)}(X_n))) - m^{(C_n)}(x_{0,j})| \geq c_{20}\cdot M_n^{-p}\Bigg\},$$

where $x_{0,j} = a_{n,j} + x_0/M_n \in A_{n,j}$. Since the $X_1, \dots, X_n, C_{n,1}, \dots, C_{n,M_n^d}$ are independent, we can reformulate the last probability as

$$\int \dots \int I_{\{\exists j \in \{1,\dots,M_n^d\} : x_1 \notin A_{n,j},\dots,x_n \notin A_{n,j}\}}$$
$$\cdot \mathbf{P}\Bigg\{|m_n(x_{0,j},(x_1,m^{(C_n)}(x_1)),\dots,(x_n,m^{(C_n)}(x_n))) - m^{(C_n)}(x_{0,j})| \geq c_{20}\cdot M_n^{-p}\Bigg\}$$
$$d\mathbf{P}_{X_n}(x_n)\dots d\mathbf{P}_{X_1}(x_1).$$

By definition of x_0 we have $m^{(C_n)}(x_{0,j}) = C_{n,j}\cdot M_n^{-p}\cdot g(x_0) = c_{20}\cdot M_n^{-p}\cdot C_{n,j}$. If a certain $A_{n,j}$ does not contain any of the $x_1, \dots, x_n$, then $m_n(x_{0,j},(x_1,m^{(C_n)}(x_1)),\dots,(x_n,m^{(C_n)}(x_n)))$ is independent of $C_{n,j}$, from which we can conclude that

$$\mathbf{P}\Bigg\{|m_n(x_{0,j},(x_1,m^{(C_n)}(x_1)),\dots,(x_n,m^{(C_n)}(x_n))) - m^{(C_n)}(x_{0,j})| \geq c_{20}\cdot M_n^{-p}\Bigg\} \geq \frac{1}{2}.$$

Summarizing the above results we see that we have shown

$$\sup_{m\in\mathcal{F}^{(p,C)}} \mathbf{P}\left\{\|m_n(\cdot,(X_1,m(X_1)),\dots,(X_n,m(X_n))) - m\|_{\infty,[0,1]^d} \geq c_{20}\cdot\left(\frac{\log n}{2\cdot n}\right)^{\frac{p}{d}}\right\}$$
$$\geq \frac{1}{2}\cdot\mathbf{P}\left\{\exists j\in\left\{1,\dots,M_n^d\right\} : X_1\notin A_{n,j},\dots,X_n\notin A_{n,j}\right\}.$$

Hence it suffices to show that

$$\liminf_{n\to\infty}\mathbf{P}\left\{\exists j\in\left\{1,\dots,M_n^d\right\} : X_1\notin A_{n,j},\dots,X_n\notin A_{n,j}\right\} > 0. \tag{14}$$

The event in (14) describes the random allocation of n balls into M_n^d urns and its probability is the classical probability of leaving at least one urn empty. We believe that its lower bound has already been computed in the literature, but since we could not find a proper reference, we provide the rigorous derivation of it below.

Since the probability in (14) is monotonically increasing in M_n, and since M_n satisfies for sufficiently large n the relation $M_n^d \geq \lfloor n/(\log n - \log\log n) \rfloor$, we can assume without loss of generality that we have $d = 1$ and $M_n = \lfloor n/(\log n - \log\log n) \rfloor$. Let C_j be the event that $A_{n,j}$ remains empty. Then we are interested in the probability $\mathbf{P}\left\{\bigcup_{j=1}^{M_n^d} C_j\right\}$. According to the inclusion-exclusion principle (formula of Sylvester-Poincaré) it can be written as

$$\mathbf{P}\left\{\bigcup_{j=1}^{M_n} C_j\right\} = \sum_{k=1}^{M_n} \sum_{\substack{I\subseteq\{1,\dots,M_n\},\\ |I|=k}} (-1)^{|I|-1}\,\mathbf{P}\left\{\bigcap_{i\in I} C_i\right\}$$

$$= \sum_{k=1}^{M_n} (-1)^{k-1} \cdot \binom{M_n}{k} \cdot \left(1 - \frac{k}{M_n}\right)^n .$$

By the tedious but not very difficult proof it is possible to show that this probability tends to $1 - \frac{1}{e}$ for n tending to infinity, which implies the assertion (cf., Appendix). □

References

[1] Beirlant, J. and Györfi, L. (1998). On the asymptotic L_2-error in partitioning regression estimation. *Journal of Statistical Planning and Inference*, **71**, pp. 93–107.

[2] Devroye, L. and Wagner, T. J. (1980). Distribution-free consistency results in nonparametric discrimination and regression function estimation. *Annals of Statistics*, **8**, pp. 231–239.

[3] Devroye, L. (1982). Necessary and sufficient conditions for the almost everywhere con-

vergence of nearest neighbor regression function estimates. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, **61**, pp. 467–481.

[4] Devroye, L. and Krzyżak, A. (1989). An equivalence theorem for L_1 convergence of the kernel regression estimate. *Journal of Statistical Planning and Inference*, **23**, pp. 71–82.

[5] Devroye, L., Györfi, L., Krzyżak, A., and Lugosi, G. (1994). On the strong universal consistency of nearest neighbor regression function estimates. *Annals of Statistics*, **22**, pp. 1371–1385.

[6] Farwig, R. (1986). Multivariate interpolation of arbitrarily spaced data by moving least squares methods. *Journal of Computational and Applied Mathematics*, **16**, pp. 79-93.

[7] Le Gia, Q. T., Narcowich, F. J., Ward, J. D., and Wendland, H. (2006). Continuous and discrete least-squares approximation by radial basis functions on spheres. *Journal of Approximation Theory*, **143**, pp. 124-133.

[8] Györfi, L. (1981). Recent results on nonparametric regression estimate and multiple classification. *Problems of Control and Information Theory*, **10**, pp. 43–52.

[9] Györfi, L., Kohler, M., Krzyżak, A. and Walk, H. (2002). *A Distribution-Free Theory of Nonparametric Regression.* Springer-Verlag, New York.

[10] Jetter, K., Buhmann, M. D., Haussmann, W., Schaback, R., and Stöckler, J. (ed.)(2006), "Topics in Multivariate Approximation and Interpolation," *Studies in Computational Mathematics*, 12, 191-230.

[11] Johnson, M. J., Shen, Z., and Xu, Y. (2009). Scattered data reconstruction by regularization in B-spline and associated wavelet spaces. *Journal of Approximation Theory*, **159**, pp. 197-223.

[12] Joldes, G. R., Chowdhury, H. A., Wittek, A., Doyle, B. and Miller, K. (2015). Modified moving least squares with polynomial bases for scattered data approximation. *Applied Mathematics and Computation*, **266**, pp. 893-902.

[13] Kohler, M. (2000). Inequalities for uniform deviations of averages from expectations with applications to nonparametric regression. *Journal of Statistical Planning and Inference*, **89**, pp. 1-23.

[14] Kohler, M. (2014). Optimal global rates of convergence for noiseless regression estimation problems with adaptively chosen design. *Journal of Multivariate Analysis*, **132**, pp. 197–208.

[15] Kohler, M. and Krzyżak, A. (2001). Nonparametric regression estimation using penalized least squares. *IEEE Transactions on Information Theory*, **47**, pp. 3054–3058.

[16] Kohler, M. and Krzyżak, A. (2013). Optimal global rates of convergence for interpolation problems with random design. *Statistics and Probability Letters*, **83**, pp. 1871–1879.

[17] Lancaster, P. and Salkauskas, K. (1981). Surfaces Generated by Moving Least Squares Methods. *Mathematics of Computation*, **37**, pp. 141-158.

[18] Lazzaro, D., and Montefusco, L. (2002), "Radial Basis Functions for the Multivariate Interpolation of Large Scattered Data Sets," *Journal of Computational and Applied Mathematics*, 140, 521-536.

[19] Levin, D. (1998). The approximation power of moving least-squares. *Mathematics of Computation*, **67**, pp. 1517-1531.

[20] Lugosi, G. and Zeger, K. (1995). Nonparametric estimation via empirical risk minimization. *IEEE Transactions on Information Theory*, **41**, pp. 677–687.

[21] Madych, W. R. and Nelson, S. A. (1992). Bounds on multivariate polynomials and exponential error estimates for multiquadric interpolation. *Journal of Approximation Theory*, **70**, pp. 94-114.

[22] Micchelli, C. A. (1986). Interpolation of scattered data: distance matrices and conditionally positive definite functions. *Constructive Approximation*, **2**, pp. 11-22.

[23] Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability and its Applications*, **9**, pp. 141–142.

[24] Nadaraya, E. A. (1970). Remarks on nonparametric estimates for density functions and regression curves. *Theory of Probability and its Applications*, **15**, pp. 134–137.

[25] Narcowich, F. J., Ward, J. D. and Wendland, H. (2006) Sobolev Error Estimates and a Bernstein Inequality for Scattered Data Interpolation via Radial Basis Functions *Constructive Approximation*, **24**, pp. 175-186.

[26] Ohtake, Y., Belyaev, A., and Seidel, H. P. (2005). 3D scattered data interpolation and approximation with multilevel compactly supported RBFs. *Graphical Models*, **67**, pp. 150-165.

[27] Ohtake, Y., Belyaev, A. and Seidel, H. P. (2006). Sparse surface reconstruction with adaptive partition of unity and radial basis functions. *Graphical Models*, **68**, 15-24.

[28] Stone, C. J. (1977). Consistent nonparametric regression. *Annals of Statistics*, **5**, pp. 595–645.

[29] Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Annals of Statistics*, **10**, pp. 1040–1053.

[30] Wahba, G. (1990). *Spline Models for Observational Data.* SIAM, Philadelphia, PA.

[31] Watson, G. S. (1964). Smooth regression analysis. *Sankhya Series A*, **26**, pp. 359–372.

[32] Wendland, H. (2001). Local polynomial reproduction and moving least squares approximation. *IMA Journal of Numerical Analysis*, **21**, pp. 285-300.

[33] Wendland, H. (2005). *Scattered data approximation.* Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, Cambridge, UK.

APPENDIX: Lower Bound for the Probability in the Proof of Theorem 2

Lemma 4. *Let* $n \in \mathbb{N}$ *and set* $M_n = \left\lfloor \frac{n}{\log n - \log\log n} \right\rfloor$. *Then*

$$\sum_{k=1}^{M_n} (-1)^{k-1} \cdot \binom{M_n}{k} \cdot \left(1 - \frac{k}{M_n}\right)^n \to 1 - \frac{1}{e} \quad (n \to \infty).$$

Proof. Throughout the proof we will apply several times the following consequence of the Lagrange formula for the remainder of a Taylor expansion: For any $x \in (0,1)$ there exists $\xi_x \in (0,x)$ such that

$$\log(1-x) = -x - \frac{1}{2 \cdot (1-\xi_x)^2} \cdot x^2.$$

In the first step of the proof we show

$$\binom{M_n}{k} \cdot \left(1 - \frac{k}{M_n}\right)^n \to \frac{1}{k!} \quad (n \to \infty) \tag{15}$$

for any $k \in \mathbb{N}_0$. Because of

$$\begin{aligned}
&\binom{M_n}{k} \cdot \left(1 - \frac{k}{M_n}\right)^n \\
&= \frac{1}{k!} \cdot M_n \cdot (M_n - 1) \cdots (M_n - k + 1) \cdot \left(\left(1 - \frac{1}{M_n}\right)^n\right)^k \cdot \left(\frac{1 - \frac{k}{M_n}}{\left(1 - \frac{1}{M_n}\right)^k}\right)^n \\
&== \frac{1}{k!} \cdot 1 \cdot \left(1 - \frac{1}{M_n}\right) \cdots \left(1 - \frac{k-1}{M_n}\right) \cdot \left(M_n \cdot \left(1 - \frac{1}{M_n}\right)^n\right)^k \cdot \left(\frac{1 - \frac{k}{M_n}}{\left(1 - \frac{1}{M_n}\right)^k}\right)^n,
\end{aligned}$$

the assertion of step 1 follows from $M_n \to \infty$ $(n \to \infty)$,

$$\left(\frac{1 - \frac{k}{M_n}}{\left(1 - \frac{1}{M_n}\right)^k}\right)^n \to 1 \quad (n \to \infty) \tag{16}$$

and

$$M_n \cdot \left(1 - \frac{1}{M_n}\right)^n \to 1 \quad (n \to \infty). \tag{17}$$

Here (16) follows from

$$\begin{aligned}
& n \cdot \left(\log\left(1 - \frac{k}{M_n}\right) - k \cdot \log\left(1 - \frac{1}{M_n}\right)\right) \\
&= n \cdot \left(-\frac{k}{M_n} - \frac{1}{2 \cdot (1 - \xi_{k/M_n})^2} \cdot \frac{k^2}{M_n^2} - k \cdot \left(-\frac{1}{M_n} - \frac{1}{2 \cdot (1 - \xi_{1/M_n})^2} \cdot \frac{1}{M_n^2}\right)\right) \\
&= \frac{n}{M_n^2} \cdot \left(\frac{k}{2 \cdot (1 - \xi_{1/M_n})^2} - \frac{k^2}{2 \cdot (1 - \xi_{k/M_n})^2}\right) \to 0 \quad (n \to \infty).
\end{aligned}$$

Furthermore, the definition of M_n implies

$$\begin{aligned}
& \log M_n + n \cdot \log\left(1 - \frac{1}{M_n}\right) \\
&= \log M_n + n \cdot \left(-\frac{1}{M_n} - \frac{1}{2 \cdot (1 - \xi_{1/M_n})^2} \cdot \frac{1}{M_n^2}\right) \\
&= \left(\log M_n - \frac{n}{M_n}\right) - \frac{1}{2 \cdot (1 - \xi_{1/M_n})^2} \cdot \frac{n}{M_n^2} \to 0 \quad (n \to \infty),
\end{aligned}$$

hence also (17) holds.

In the second step of the proof we show

$$\left|(-1)^{k-1} \cdot \binom{M_n}{k} \cdot \left(1 - \frac{k}{M_n}\right)^n\right| \leq \frac{2^k}{k!} \quad \text{for all } k \in \{1, \ldots, M_n - 1\} \tag{18}$$

for n sufficiently large.

Since

$$\log\left(1 - \frac{k}{M_n}\right) = \sum_{l=1}^{\infty} \frac{-1}{l} \cdot \left(\frac{k}{M_n}\right)^l \leq k \cdot \sum_{l=1}^{\infty} \frac{-1}{l} \cdot \left(\frac{1}{M_n}\right)^l = k \cdot \log\left(1 - \frac{1}{M_n}\right),$$

we have

$$\left|(-1)^{k-1} \cdot \binom{M_n}{k} \cdot \left(1 - \frac{k}{M_n}\right)^n\right| \quad \leq \quad \binom{M_n}{k} \cdot \left(1 - \frac{1}{M_n}\right)^{n \cdot k}$$

$$\leq \frac{1}{k!} \cdot \left(M_n \cdot \left(1 - \frac{1}{M_n}\right)^n \right)^k \leq \frac{2^k}{k!},$$

for n sufficiently large, where the last inequality follows from (17).

In the third step of the proof we show the assertion. Here we apply the dominated convergence theorem together with (15) and (18) and conclude

$$\sum_{k=1}^{M_n} (-1)^{k-1} \cdot \binom{M_n}{k} \cdot \left(1 - \frac{k}{M_n}\right)^n$$

$$= \sum_{k=1}^{\infty} (-1)^{k-1} \cdot \binom{M_n}{k} \cdot \left(1 - \frac{k}{M_n}\right)^n \cdot I_{\{k \leq M_n - 1\}} \to \sum_{k=1}^{\infty} \frac{(-1)^{k-1}}{k!} = 1 - \frac{1}{e}$$

for $n \to \infty$. □