# Nonparametric estimation of a function from noiseless observations at random points

Benedikt Bauer, Luc Devroye, Michael Kohler, Adam Krzyżak,
Harro Walk

PII: S0047-259X(17)30338-X
DOI: http://dx.doi.org/10.1016/j.jmva.2017.05.010
Reference: YJMVA 4255

To appear in: *Journal of Multivariate Analysis*

Received date : 12 August 2016

Please cite this article as: B. Bauer, L. Devroye, M. Kohler, A. Krzyżak, H. Walk, Nonparametric estimation of a function from noiseless observations at random points, *Journal of Multivariate Analysis* (2017), http://dx.doi.org/10.1016/j.jmva.2017.05.010

# Nonparametric estimation of a function from noiseless observations at random points[*]

Benedikt Bauer[a], Luc Devroye[b], Michael Kohler[a], Adam Krzyżak[c,†] and Harro Walk[d]

[a] Fachbereich Mathematik, Technische Universität Darmstadt, Schlossgartenstraße 7, 64289 Darmstadt, Germany
Email: bbauer@mathematik.tu-darmstadt.de, kohler@mathematik.tu-darmstadt.de

[b] School of Computer Science, McGill University, 3450, rue University, Montréal (QC), Canada H3A 2K6
Email: lucdevroye@gmail.com

[c] Department of Computer Science and Software Engineering, Concordia University,
1455, boul. de Maisonneuve ouest, Montréal (QC) Canada H3G 1M8
Email: krzyzak@cs.concordia.ca

[d] Fachbereich Mathematik, Universität Stuttgart, Pfaffenwaldring 57, 70569 Stuttgart, Germany
Email: walk@mathematik.uni-stuttgart.de

May 27, 2017

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 whose 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 $\{\ln(n)/n\}^{p/d}$ for arbitrary $p$ and $d$ and that this rate of convergence is optimal in minimax sense.

*AMS classification:* Primary 62G05; secondary 62G20.

*Key words and phrases:* Multivariate scattered data approximation, rate of convergence, supremum norm error.

## 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 $(x_1, y_1), \ldots, (x_n, y_n) \in [0, 1]^d \times \mathbb{R}$ 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 [6, 11, 16, 18, 32, 33], schemes based on radial basis functions or constant functions on spheres [10, 17, 24–26], multiquadric interpolants [21] and the smoothing spline approach. The latter one can be posed as the regularized least squares problem where one minimizes the criterion

$$\sum_{i=1}^{n} \{m(x_i) - y_i\}^2 + \lambda \, \|m\|_H^2$$

[*] Running title: *Nonparametric estimation of a function*

[†] Corresponding author. Tel: +1–514–848–2424 ext. 3007, Fax: +1–514–848–2830

over a class of functions $H$. The classes of functions include Beppo-Levi Space [10] and Reproducing Kernel Hilbert Space [7]. In the moving least squares approach we seek function $m^*$ which is a solution of the minimization problem

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

where $P$ is a finite-dimensional subspace (usually spanned by polynomials) of a space of continuous functions on a compact set $\Omega$. 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 [32]. For the rate of approximation define the separation distance $q_X$ and the mesh norm $h_{X,\Omega}$ as follows:

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

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

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

holds on the data set. Then under the condition that $\Omega$ is compact and satisfies the so-called cone condition we get for $f \in C^p(\Omega)$ the approximation bound $\|m - m^*\|_{\infty,\Omega} \le c_2 \, h^p_{X,\Omega}$; see, e.g., [32, 33]. Hence if $x_1, \dots, x_n$ are scattered approximately evenly in $[0,1]^d$, we get

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

The approximation error bounds for the radial basis function interpolations may be found in [33] and [20].

### 1.2. The problem studied in this paper

In practice it is not clear, especially in high dimensions, at which locations a 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 $X, X_1, \dots, X_n$ 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 $\mathcal{D}_n = \{(X_1, m(X_1)), \dots, (X_n, m(X_n))\}$, 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 (see, e.g., [9], 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

$$\Pr\{X \in S_r(x)\} > c_4 \, r^d$$

does hold for all $x \in [0,1]^d$ and all $0 < r \le 1$, 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 a spline estimate $m_n = m_n(\cdot, \mathcal{D}_n)$ such that

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

where we write $Z_n = O_{\mathbf{P}}(Y_n)$ if the nonnegative random variables $Z_n$ and $Y_n$ satisfy $\lim_{c\to\infty} \limsup_{n\to\infty} \Pr(Z_n > c\, Y_n) = 0$. 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 (see, e.g., [4, 5, 22, 23, 28, 29, 31], partitioning regression estimates (see, e.g., [1, 8]), nearest neighbor regression estimates (see, e.g., [2, 3]), least squares estimates (see, e.g., [12, 19]) and smoothing spline estimates (see, e.g., [14, 30]).

Minimax rate of convergence results for the global errors of such estimates have been derived in [29]. 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 $n^{-2p/(2p+d)}$.

In the setting of fixed design regression estimation it was analyzed in [13] how the above rate of convergence changes if there is no noise in the dependent variable. The main results there are that for suitably defined spline estimates the supremum norm error converges to zero with the rate $n^{-p/d}$ (which corresponds to the bound (3)) and that this rate of convergence is optimal in some minimax sense.

For the problem studied in this article it was shown in [15] that that the expected $L_1$-error of a nearest-neighbor estimate achieves the rate of convergence $n^{-p/d}$ 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 similar to (3) lead to rates (4) or similar, however it is not clear how one can construct an estimate for random scattered data achieving the rate given by (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 $\mathbb{R}$, respectively. 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 $\alpha_1 + \cdots + \alpha_d = k$ the partial derivative $\partial^k f/(\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d})$ exists and satisfies

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

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 nonnegative constant.

### 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 reliable. 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 require many 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, we 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 $u_j = j/K$ for indices $j \in \{-M, \ldots, K + M\}$. The (univariate)* B-splines *$B_{j,\ell} : \mathbb{R} \to \mathbb{R}$ of degree $\ell$ are recursively defined by*

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

*for $j \in \{-M, \ldots, K + M - 1\}$ and*

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

*for $j \in \{-M, \ldots, K + M - l - 2\}$ and $\ell \in \{0, \ldots, M - 1\}$.*

*The sequence $(u_j)_{j \in \{-M,\ldots,K+M\}}$ 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 $\mathbf{j} = (j_1, \ldots, j_d)$ with $\mathbf{j} \in \{-M, \ldots, K + M\}^d$, the* tensor product B-spline *$B_{\mathbf{j},M} : \mathbb{R}^d \to \mathbb{R}$ is defined by*

$$B_{\mathbf{j},M}(x) = B_{j_1,M}(x^{(1)}) \cdots B_{j_d,M}(x^{(d)}).$$

For $M \in \mathbb{N}$ and $K = K_n = \lceil c_5 (n/\ln n)^{1/d} \rceil$ 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, for all $x \in [0, 1]^d$, by

$$m_n(x) = \sum_{\mathbf{j} \in \{-M,\ldots,K_n-1\}^d} \hat{c}_{\mathbf{j}} \, B_{\mathbf{j},M}(x) \tag{5}$$

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 [13] that it is possible to choose these coefficients such that $|m_n(x) - m(x)| \leq c_6 \, K_n^{-p}$ holds for a constant $c_6 > 0$ depending only on $d, M, p$ and $C$. So if we set $\delta_n = c_7 \, K_n^{-p}$ for a suitably chosen $c_7 > 0$, we will find coefficients $\hat{c}_{\mathbf{j}}$ such that the following $n$ inequalities are satisfied:

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

Since $c_7 \, K_n^{-p}$ (especially $p$) is usually not known for the function $m$, we choose $\delta_n$ adaptively in the following way to compute our estimate by (6):

$$\delta_n = \min\{2^\ell : \ell \in \{-n, \ldots, n\} \text{ and system (6) is solvable using } 2^\ell \text{ as right-hand side}\}.$$

If the above set is empty, we define $m_n = 0$.

For $n$ sufficiently large and $m$ $(p, C)$-smooth, the set $\delta_n$ is chosen from (and the corresponding solution space of system (6)) must be non-empty because of the above-mentioned result in Kohler [13] and the fact that $2^n$ becomes larger than $c_7 \, K_n^{-p}$ for increasing $n$. Linear programming can be used to compute the coefficients $\hat{c}_{\mathbf{j}} \in \mathbb{R}$ in practice.

## 3. Main results

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

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

$$\Pr\{X \in S_r(x)\} > c_8\, r^d$$

*holds for all $x \in [0,1]^d$ and all $r \in (0,1]$. 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 choose $K_n$ and $\delta_n$ as in Section 2.2. Let $m_n$ be defined by (5) and (6). Then for $c_5 > 0$ sufficiently small we have*

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

**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 $c_{11} > 0$

$$\limsup_{n\to\infty} \sup_{m\in\mathcal{F}^{(p,C)}} \Pr[\|m_n - m\|_{\infty,[0,1]^d} \geq c_{11}\{\ln(n)/n\}^{p/d}] = 0,$$

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

**Remark 2.** The conditions of Theorem 1 require the probability of any ball whose center lies within the unit cube to be bounded from below in a certain way. On the one hand, this assumption could be even further weakened because (as a referee pointed out) it must be satisfied only for every radius greater than a small constant times $\{\ln(n)/n\}^{p/d}$. On the other hand, it is an interesting question, which distributions of $X$ actually satisfy the requirements of Theorem 1. A more specific (and easily understandable) type of random variables complying with this condition are those, which have a density bounded away from zero on the unit cube as stated in the Abstract of this article.

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*

$$\liminf_{n\to\infty} \inf_{m_n} \sup_{m\in\mathcal{F}^{(p,C)}} \Pr[\|m_n - m\|_{\infty,[0,1]^d} \geq c_{12}\,\{\ln(n)/n\}^{p/d}] > 0$$

*holds, where the infimum is computed with respect to all estimates $m_n$ depending on $(X_1, m(X_1)), \dots, (X_n, m(X_n))$.*

**Remark 3.** Although the rate of convergence deduced in Theorem 1 is optimal according to Theorem 2, it suffers from the so-called curse of dimensionality. This means that the rate becomes worse (for fixed smoothness) if the dimension $d$ is increased.

## 4. Application to simulated data

In this section we apply the estimate developed in the previous sections to simulated data and compare the results with conventional estimates using the statistics package R. Whereas Theorem 1 and Theorem 2 revealed the optimal asymptotic behavior of our new estimate, it is not clear how the estimate behaves in case of small sample sizes. This will be examined in the following.

For this purpose, we consider three competitive approaches. The first one is interpolation with radial basis functions presented in [17], where authors use Wendland's compactly supported radial basis function $\phi(r) = (1-r)_+^6\,(35r^2 + 18r + 3)$. 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 [11], where we scale the radius of influence they

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)** |

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 leads 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 $\delta_n$ is chosen as the smallest possible value in $\{2^i/n : i = -50, \ldots, 30\}$ 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 $m_i : [0,1]^2 \to \mathbb{R}$, $i \in \{1, \ldots, 6\}$, with different degrees of $(p, C)$-smoothness as illustrated in Figure 1 and defined as follows:

$$\begin{aligned} m_1(x_1, x_2) &= 3x_1^2 x_2 - x_2^3, \\ m_2(x_1, x_2) &= 2\exp\{-5\,(x_1 - 0.7)^2\} - \exp\{-5\,(x_1 - 0.4)^2\} - 3x_2 + 5, \end{aligned}$$

$$\begin{aligned}
m_3(x_1,x_2) &= 1/(x_1+x_2^3+0.5)\\
m_4(x_1,x_2) &= \exp[-3\{(x_1-0.75)^2+(x_2-0.75)^2\}],\\
m_5(x_1,x_2) &= \sin(2\pi x_1)\cos(\pi x_2),\\
m_6(x_1,x_2) &= \min(1-x_2, 2x_1-0.5).
\end{aligned}$$

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 for increasing sample size. All of these results support the following conclusion regarding the small sample size behavior of our new estimate: It outperforms other approaches in case of smooth functions, but it faces problems in case of non-smooth functions.

## 5. Proofs

### *5.1. Proof of Theorem 1*

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

**Lemma 1.** *Let $\Pi$ 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 $\Pi$ with dimension $\dim\mathcal{P}$. For Lebesgue almost any $(x_1^{(1)},\dots,x_1^{(d)},\dots,x_{\dim\mathcal{P}}^{(1)},\dots,x_{\dim\mathcal{P}}^{(d)}) \in \mathbb{R}^{d\,\dim\mathcal{P}}$ and any $y_1,\dots,y_{\dim\mathcal{P}} \in \mathbb{R}$, there is a unique $p \in \mathcal{P}$ which fulfills $p(x_i^{(1)},\dots,x_i^{(d)}) = y_i$ for all $i \in \{1,\dots,\dim\mathcal{P}\}$.*

**Proof.** The assertion of the lemma follows immediately from Proposition 4 in [27]. □

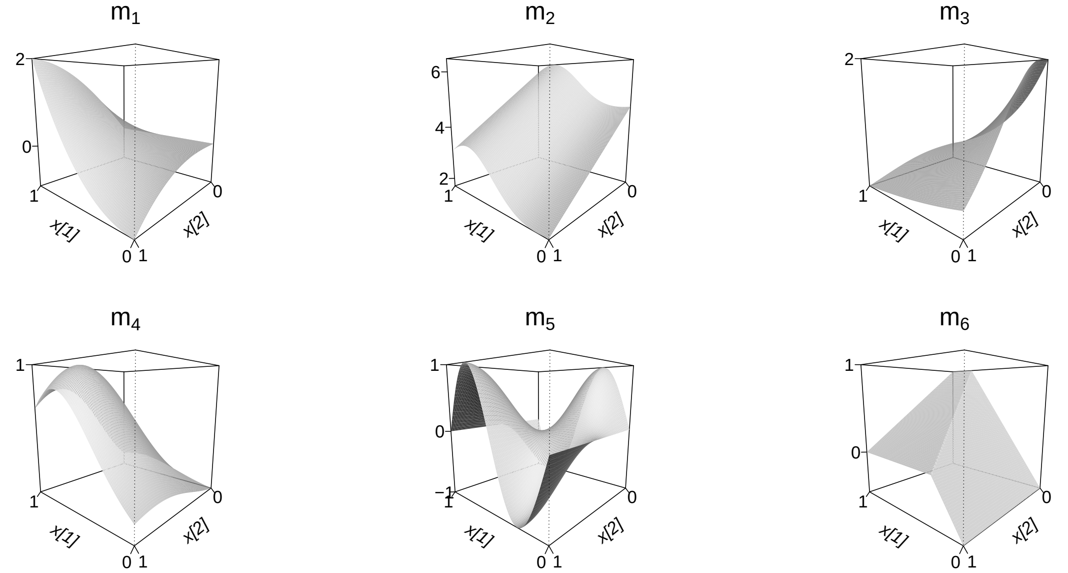

Figure 1: Behavior of the test functions $m_1, m_2, m_3, m_4, m_5, m_6$.

**Lemma 2.** *Assume that the distribution of the iid random variables $X, X_1, \dots, X_n$ satisfies $\Pr\{X \in S_\varepsilon(x)\} \geq c_{13}\,\varepsilon^d$ 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 $\varepsilon$. Let $K_n = \lceil c_9\,\{n/\ln(n)\}^{1/d}\rceil$. Let $B_1, \dots, B_{K_n^d}$ be $K_n^d$ balls with radius $c_{14}\,1/K_n$, whose centers lie in $[0,1]^d$. For any $r > 0$, there is a sufficiently small $c_9 \equiv c_9(r, c_{13}, c_{14}) > 0$, such that*

$$\lim_{n\to\infty} \Pr\{\forall_{j\in\{1,\dots,K_n^d\}}\ \exists_{i\in\{1,\dots,\lfloor n/r\rfloor\}} : X_i \in B_j\} = 1.$$

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

$$\begin{aligned}
\Pr\left\{\exists_{j\in\{1,\dots,K_n^d\}} : X_1, \dots, X_{\lfloor n/r\rfloor} \notin B_j\right\} &\leq \sum_{j\in\{1,\dots,K_n^d\}} \{1 - \Pr(X \in B_j)\}^{\lfloor n/r\rfloor} \\
&\leq K_n^d \max_{j\in\{1,\dots,K_n^d\}} \{1 - \Pr(X \in B_j)\}^{\lfloor n/r\rfloor} \\
&\leq K_n^d\,(1 - c_{13}c_{14}^d/K_n^d)^{\lfloor n/r\rfloor} \\
&\leq K_n^d \exp\left(-c_{13}\,c_{14}^d\,\frac{n}{2\,r\,K_n^d}\right) \\
&\leq 2^d\,c_9^d\,\frac{n}{\ln(n)} \exp\left\{-c_{13}\,c_{14}^d\,\frac{\ln(n)}{2^{d+1}\,r\,c_9^d}\right\}
\end{aligned}$$

For sufficiently small $c_9$, the right-hand side of the inequality above tends to zero as $n \to \infty$. □

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

$$\lim_{n\to\infty} \Pr\{\forall_{j\in\{1,\dots,r\,K_n^d\}}\ \exists_{i\in\{1,\dots,n\}} : X_i \in B_j\} = 1.$$

**Proof.** At first, we note that

$$\Pr\{\forall_{j\in\{1,\dots,r\cdot K_n^d\}}\ \exists_{i\in\{1,\dots,n\}} : X_i \in B_j\} \geq \prod_{k=1}^{r} \Pr\{\forall_{j\in\{(k-1)\cdot K_n^d+1,\dots,k\cdot K_n^d\}}\ \exists_{i\in\{(k-1)\cdot\lfloor n/r\rfloor+1,\dots,k\cdot\lfloor n/r\rfloor\}} : X_i \in B_j\}.$$

The assertion follows from the application of Lemma 2 to 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 [13] which ensures that

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

fulfills $|\bar{m}_n(x) - m(x)| \leq c_{15}\,K_n^{-p}$ for all $x \in [0,1]^d$ and a constant $c_{15} > 0$. Consequently, for $n$ sufficiently large, a right-hand side of the type $2^\ell$ with $\ell \in \{-n, \dots, n\}$ which makes system (6) solvable exists and the smallest value of this type is smaller than $2\,c_{15}\,K_n^{-p}$. Then the bound $\delta_n \leq c_{10}\,K_n^{-p}$ holds. These considerations imply

$$|m_n(X_i) - \bar{m}_n(X_i)| \leq |m_n(X_i) - m(X_i)| + |\bar{m}_n(X_i) - m(X_i)| \leq \delta_n + c_{15}\,K_n^{-p} \leq c_{16}\,\{\ln(n)/n\}^{p/d} \tag{7}$$

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

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

holds, so we can conclude

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

for all $x \in [0,1]^d$, because the B-splines are non-negative and sum up to 1 (see, e.g., Lemma 15.2 in [9]). Combining this with the previous bounds we get

$$|m_n(x) - m(x)| \leq |m_n(x) - \bar{m}_n(x)| + |\bar{m}_n(x) - m(x)| \leq \max_{\mathbf{j}\in\{-M,\ldots,K-1\}^d} |z_{\mathbf{j}}| + c_{17}\, t\{\ln(n)/n\}^{p/d}$$

and thus it suffices to show that $\max_{\mathbf{j}\in\{-M,\ldots,K_n-1\}^d} |z_{\mathbf{j}}| \leq c_{18}\,\{\ln(n)/n\}^{p/d}$ for a certain $c_{18} > 0$ outside of an event, whose probability tends to zero for $n \to \infty$.

By (7) our estimate fulfills, for each $i \in \{1,\ldots,n\}$,

$$\sum_{\mathbf{j}\in\{-M,\ldots,K_n-1\}^d} z_{\mathbf{j}}\, B_{\mathbf{j},M}\,(X_i) = \varepsilon(i) \tag{8}$$

for an adequately chosen

$$\varepsilon(i) \in \left[-c_{16}\,\{\ln(n)/n\}^{p/d}, c_{16}\,\{\ln(n)/n\}^{p/d}\right]. \tag{9}$$

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

$$\forall_{i\in\{1,\ldots,d\}} \quad j_i - M \leq k_i \leq j_i.$$

If we put $(M+1)^d$ different values $x_1,\ldots,x_{(M+1)^d} \in A_{\mathbf{j}}$ in equations of the type (8), this leads to the linear system of equations

$$\forall_{i\in\{1,\ldots,(M+1)^d\}} \quad \sum_{\mathbf{k}\in\mathcal{S}_{\mathbf{j}}} z_{\mathbf{k}}\, B_{\mathbf{k},M}\,(x_i) = \varepsilon(i), \tag{10}$$

because the rest of the B-spline terms vanish on $A_{\mathbf{j}}$. We will abbreviate (10) in matrix notation by $\mathbf{B_j}\,\mathbf{z_j} = \varepsilon_{\mathbf{j}}$. Because the remaining B-splines are polynomials on this set, (10) equals a polynomial interpolation problem on $A_{\mathbf{j}}$.

Let $\bar{B}_{k,M}$ be the univariate B-spline of degree $M$ with knot sequence $\bar{u}_k = k$ ($k \in \mathbb{Z}$) and support $[k, k+M+1]$, and set

$$\bar{B}_{\mathbf{k},M}\,(x) = B_{k_1,M}(x^{(1)}) \cdots B_{k_d,M}(x^{(d)})$$

for $\mathbf{k} = (k_1,\ldots,k_d)$. Then it is easy to see that

$$B_{k,M}(x) = \bar{B}_{k,M}\{K_n\ (x - k/K_n)\} = \bar{B}_{k,M}\,(K_n\, x - k)\,,$$

and consequently we can consider the polynomial interpolation problem $\mathbf{B_j}\,\mathbf{z_j} = \varepsilon_{\mathbf{j}}$ on $(0,1)^d$ rather than on $A_{\mathbf{j}}$ and with B-splines $\bar{B}_{\mathbf{j},M}$, which are based on the nodes with node distance 1, rather than with B-splines which use node distance $1/K_n$.

Due to the local linear independence of the B-splines (see Lemma 14.5 in [9]) 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 continuous function of 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 argument 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 $c_{19}/K_n$ 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 (8) form a system like (10) which can be solved by Cramer's rule in the form of

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

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

$$|z_{\mathbf{k}}| = \frac{|\det\{\mathbf{B_j}(\mathbf{k}, \varepsilon_{\mathbf{j}})\}|}{|\det(\mathbf{B_j})|} \le \frac{c_{20}}{c_{\min}} \max_{i\in\{1,\dots,(M+1)^d\}} |\varepsilon(i)| \le c_{18}\,\{\ln(n)/n\}^{p/d} \tag{11}$$

because of (9). Since the above works for all of the $A_{\mathbf{j}}$ simultaneously (see Lemma 3), every $z_{\mathbf{j}}$ in (8) can be bounded by (11), and this implies the assertion. □

### 5.2. *Proof of Theorem 2*

Set $M_n = \lfloor \{2\,n/\ln(n)\}^{1/d} \rfloor$ and let $\left\{A_{n,j}\right\}_{j=1,\dots,M_n^d}$ be a partition of $[0,1]^d$ into cubes of side length $1/M_n$. 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

$$\mathrm{supp(g)} \subseteq (-1/2, 1/2)^d$$

and reaching a certain constant $c_{21} > 0$ on its support, i.e., satisfying $c_{21} = \sup_{x\in\mathbb{R}^d} g(x) = g(x_0) > 0$ for some $x_0 \in (-1/2, 1/2)^d$ . For $j \in \{1, \dots, M_n^d\}$ let $a_{n,j}$ be the center of $A_{n,j}$ and set $g_{n,j}(x) = M_n^{-p}\, g\{M_n(x - a_{n,j})\}$. We define $m^{(c_n)} : \mathbb{R}^d \to \mathbb{R}$ by

$$m^{(c_n)} = \sum_{j=1}^{M_n^d} c_{n,j}\, g_{n,j}(x),$$

where $c_n = (c_{n,j})_{j=1,\dots,M_n^d} \in \{-1,1\}^{M_n^d}$.

The functions $m^{(c_n)}$ are $(p,C)$-smooth for all $c_n \in \{-1,1\}^{M_n^d}$ (see, e.g., [9], proof of Theorem 3.2), hence we have

$$\{m^{(c_n)} \,:\, c_n \in \{-1,1\}^{M_n^d}\} \subseteq \mathcal{F}^{(p,C)}. \tag{12}$$

Randomizing the coefficients of this type of functions we introduce random variables $C_{n,1}, \dots, C_{n,M_n^d}$ which are independent from each other and from $X_1, \dots, X_n$, such that

$$\Pr\{C_{n,k} = -1\} = \Pr\{C_{n,k} = 1\} = 1/2$$

for all $k \in \{1, \dots, M_n^d\}$, and we set $C_n = (C_{n,1}, \dots, C_{n,M_n^d})$. Using the relation $M_n \le \{2\,n/\ln(n)\}^{1/d}$, (12) allows the following bounding for an arbitrary estimate $m_n$:

$$\begin{aligned}
&\sup_{m\in\mathcal{F}^{(p,C)}} \Pr\Big[\|m_n(\cdot, (X_1, m(X_1)), \dots, (X_n, m(X_n))) - m\|_{\infty,[0,1]^d} \ge c_{21}\ \{\ln(n)/(2n)\}^{p/d}\Big] \\
&\qquad \ge \sup_{c_n\in\{-1,1\}^{M_n^d}} \Pr\Big\{\|m_n(\cdot, (X_1, m^{(c_n)}(X_1)), \dots, (X_n, m^{(c_n)}(X_n))) - m^{(c_n)}\|_{\infty,[0,1]^d} \ge c_{21}\, M_n^{-p}\Big\} \\
&\qquad \ge \Pr\Big\{\|m_n(\cdot, (X_1, m^{(C_n)}(X_1)), \dots, (X_n, m^{(C_n)}(X_n))) - m^{(C_n)}\|_{\infty,[0,1]^d} \ge c_{21}\, M_n^{-p}\Big\} \\
&\qquad \ge \Pr\Big\{\exists_{j\in\{1,\dots,M_n^d\}} : X_1 \notin A_{n,j}, \dots, X_n \notin A_{n,j} \text{ and} \\
&\qquad\qquad |m_n(x_{0,j}, (X_1, m^{(C_n)}(X_1)), \dots, (X_n, m^{(C_n)}(X_n))) - m^{(C_n)}(x_{0,j})| \ge c_{21}\, M_n^{-p}\Big\},
\end{aligned}$$

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

$$\begin{aligned}
&\int \dots \int \mathbf{1}\{\exists_{j\in\{1,\dots,M_n^d\}} \,:\, x_1 \notin A_{n,j}, \dots, x_n \notin A_{n,j}\} \\
&\qquad \times \Pr\Big\{|m_n(x_{0,j}, (x_1, m^{(C_n)}(x_1)), \dots, (x_n, m^{(C_n)}(x_n))) - m^{(C_n)}(x_{0,j})| \ge c_{21} M_n^{-p}\Big\}\mu\,(dx_n)\cdots\mu\,(dx_1)\,,
\end{aligned}$$

where $\mu$ denotes the distribution of $X$. By definition of $x_0$ we have $m^{(C_n)}(x_{0,j}) = C_{n,j}\, M_n^{-p}\, g(x_0) = c_{21}\, M_n^{-p}\, C_{n,j}$. If a certain $A_{n,j}$ does not contain any of the $x_1, \ldots, x_n$, then $m_n(x_{0,j}, (x_1, m^{(C_n)}(x_1)), \ldots, (x_n, m^{(C_n)}(x_n)))$ is independent of $C_{n,j}$, from which we can conclude that

$$\Pr\Big\{|m_n(x_{0,j}, (x_1, m^{(C_n)}(x_1)), \ldots, (x_n, m^{(C_n)}(x_n))) - m^{(C_n)}(x_{0,j})| \geq c_{21}\, M_n^{-p}\Big\} \geq 1/2.$$

Summarizing the above results we see that we have shown

$$\sup_{m\in\mathcal{F}^{(p,C)}} \Pr\Big[\|m_n(\cdot, (X_1, m(X_1)), \ldots, (X_n, m(X_n))) - m\|_{\infty,[0,1]^d} \geq c_{21}\{\ln(n)/(2n)\}^{p/d}\Big]$$
$$\geq \frac{1}{2}\, \Pr\{\exists_{j\in\{1,\ldots,M_n^d\}} : X_1 \notin A_{n,j}, \ldots, X_n \notin A_{n,j}\}.$$

Hence it suffices to show that

$$\liminf_{n\to\infty} \Pr\left\{\exists_{j\in\{1,\ldots,M_n^d\}} : X_1 \notin A_{n,j}, \ldots, X_n \notin A_{n,j}\right\} > 0. \tag{13}$$

The event in (13) 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 (13) is monotonically increasing in $M_n$, and since $M_n$ satisfies for sufficiently large $n$ the relation $M_n^d \geq \lfloor n/\{\ln(n) - \ln\ln(n)\}\rfloor$, we can assume without loss of generality that we have $d = 1$ and $M_n = \lfloor n/\{\ln(n) - \ln\ln(n)\}\rfloor$. Let $C_j$ be the event that $A_{n,j}$ remains empty. Then we are interested in the probability

$$\Pr\left(\bigcup_{j=1}^{M_n} C_j\right).$$

According to the inclusion-exclusion principle (the Sylvester–Poincaré formula), it can be written as

$$\Pr\left(\bigcup_{j=1}^{M_n} C_j\right) = \sum_{k=1}^{M_n} \sum_{\substack{I\subseteq\{1,\ldots,M_n\},\\ |I|=k}} (-1)^{|I|-1} \Pr\left(\bigcap_{i\in I} C_i\right) = \sum_{k=1}^{M_n} (-1)^{k-1} \binom{M_n}{k} \left(1 - \frac{k}{M_n}\right)^n.$$

By a tedious but not very difficult argument, it is possible to show that this probability tends to $1 - 1/e$ as $n \to \infty$, which implies the assertion (see Appendix). □

**Acknowledgments.** The first and third authors would like to thank the Deutsche Forschungsgemeinschaft (DFG) for funding this project within the Collaborative Research Centre 805. The second and fourth authors gratefully acknowledge support from the Natural Sciences and Engineering Research Council of Canada under Grants A3456 and RGPIN–2015–06412, respectively. We gratefully acknowledge the constructive comments of two anonymous referees which helped improving the manuscript and we thank the Editor-in-Chief for carefully reading the manuscript and pointing out various typos and inaccuracies.

## Appendix: Lower bound on the probability in the proof of Theorem 2

**Lemma 4.** *Let $n \in \mathbb{N}$ and set $M_n = \lfloor n/[\ln(n) - \ln\{\ln(n)\}]\rfloor$. Then*

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

**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

$$\ln(1 - x) = -x - \frac{1}{2(1 - \xi_x)^2}\, x^2.$$

*In the first step of the proof*, we show that, for any $k \in \mathbb{N}_0$,

$$\lim_{n\to\infty} \binom{M_n}{k} \left(1 - \frac{k}{M_n}\right)^n = \frac{1}{k!}\,. \tag{14}$$

Because of

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

the assertion of Step 1 follows from the fact that $M_n \to \infty$ as $n \to \infty$, which implies in turn that,

$$\lim_{n\to\infty} \left\{\frac{1-k/M_n}{(1-1/M_n)^k}\right\}^n = 1 \tag{15}$$

and

$$\lim_{n\to\infty} M_n\,(1 - \;1/M_n)^n = 1. \tag{16}$$

Here (15) follows from the fact that, as $n \to \infty$,

$$\begin{aligned} n\left\{\ln\left(1-\frac{k}{M_n}\right) - k\ln\left(1-\frac{1}{M_n}\right)\right\} &= n\left[-\frac{k}{M_n} - \frac{1}{2(1-\xi_{k/M_n})^2}\frac{k^2}{M_n^2} - k\left\{-\frac{1}{M_n} - \frac{1}{2(1-\xi_{1/M_n})^2}\frac{1}{M_n^2}\right\}\right] \\ &= \frac{n}{M_n^2}\left\{\frac{k}{2(1-\xi_{1/M_n})^2} - \frac{k^2}{2(1-\xi_{k/M_n})^2}\right\} \to 0. \end{aligned}$$

Furthermore, the definition of $M_n$ implies that, as $n \to \infty$,

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

hence also (16) holds.

*In the second step of the proof* we show that, for all $k \in \{1, \ldots, M_n - 1\}$,

$$\left|(-1)^{k-1}\binom{M_n}{k}\left(1-\frac{k}{M_n}\right)^n\right| \le \frac{2^k}{k!} \tag{17}$$

for $n$ sufficiently large. Since

$$\ln\left(1-\frac{k}{M_n}\right) = \sum_{\ell=1}^{\infty}\frac{-1}{\ell}\left(\frac{k}{M_n}\right)^{\ell} \le k\sum_{\ell=1}^{\infty}\frac{-1}{\ell}\left(\frac{1}{M_n}\right)^{\ell} = k\,\ln\left(1-\frac{1}{M_n}\right),$$

we have

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

for $n$ sufficiently large, where the last inequality follows from (16).

*In the third step of the proof*, we show the assertion. Here we apply the dominated convergence theorem together with (14) and (17) and conclude

$$\sum_{k=1}^{M_n}(-1)^{k-1}\binom{M_n}{k}\left(1-\frac{k}{M_n}\right)^n = \sum_{k=1}^{\infty}(-1)^{k-1}\binom{M_n}{k}\left(1-\frac{k}{M_n}\right)^n \mathbf{1}(k \le M_n - 1),$$

which converges to $1 - 1/e$ as $n \to \infty$. □

## References

[1] J. Beirlant, L. Györfi, On the asymptotic $L_2$-error in partitioning regression estimation, J. Stat. Plan. Inf. 71 (1998) 93–107.

[2] L. Devroye, Necessary and sufficient conditions for the almost everywhere convergence of nearest neighbor regression function estimates, Z. Wahrscheinlichkeitstheorie verw. Geb. 61 (1982) 467–481.

[3] L. Devroye, L. Györfi, A. Krzyżak, G. Lugosi, On the strong universal consistency of nearest neighbor regression function estimates, Ann. Statist. 22 (1994) 1371–1385.

[4] L. Devroye, A. Krzyżak, An equivalence theorem for $L_1$ convergence of the kernel regression estimate, J. Stat. Plan. Inf. 23 (1989) 71–82.

[5] L. Devroye, T.J. Wagner, Distribution-free consistency results in nonparametric discrimination and regression function estimation, Ann. Statist. 8 (1980) 231–239.

[6] R. Farwig, Multivariate interpolation of arbitrarily spaced data by moving least squares methods, J. Comput. Appl. Math.16 (1986) 79–93.

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

[8] L. Györfi, Recent results on nonparametric regression estimate and multiple classification, Prob. Control Inf. Theory 10 (1981) 43–52.

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

[10] M.J. Johnson, Z. Shen, Y. Xu, Scattered data reconstruction by regularization in B-spline and associated wavelet spaces, J. Approx. Theory 159 (2009) 197–223.

[11] G.R. Joldes, H.A. Chowdhury, A. Wittek, B. Doyle, K. Miller, Modified moving least squares with polynomial bases for scattered data approximation, Appl. Math. Comput. 266 (2015) 893–902.

[12] M. Kohler, Inequalities for uniform deviations of averages from expectations with application to nonparametric regression, J. Stat. Plan. Inf. 89 (2000) 1–23.

[13] M. Kohler, Optimal global rates of convergence for noiseless regression estimation problems with adaptively chosen design, J. Multivariate Anal. 132 (2014) 197–208.

[14] M. Kohler, A. Krzyżak, Nonparametric regression estimation using penalized least squares, IEEE Trans. Inf. Theory 47 (2001) 3054–3058.

[15] M. Kohler, A. Krzyżak, Optimal global rates of convergence for interpolation problems with random design, Stat. Probab. Lett. 83 (2013) 1871–1879.

[16] P. Lancaster, K. Salkauskas, Surfaces generated by moving least squares methods, Math. Comput. 37 (1981) 141–158.

[17] D. Lazzaro, L. Montefusco, Radial basis functions for the multivariate interpolation of large scattered data sets, J. Comput. Appl. Math. 140 (2002) 521–536.

[18] D. Levin, The approximation power of moving least-squares, Math. Comput. 67 (1998) 1517–1531.

[19] G. Lugosi, K. Zeger, Nonparametric estimation via empirical risk minimization, IEEE Trans. Inf. Theory 41 (1995) 677–687.

[20] W.R. Madych, S.A. Nelson, Bounds on multivariate polynomials and exponential error estimates for multiquadric interpolation, J. Approx. Theory 70 (1992) 94–114.

[21] C.A. Micchelli, Interpolation of scattered data: Distance matrices and conditionally positive definite functions, Constr. Approx. 2 (1986) 11–22.

[22] E.A. Nadaraya, On estimating regression, Th. Probab. Appl. 9 (1964) 141–142.

[23] E.A. Nadaraya, Remarks on nonparametric estimates for density functions and regression curves, Th. Probab. Appl. 15 (1970) 134–137.

[24] F.J. Narcowich, J.D. Ward, H. Wendland, Sobolev error estimates and a Bernstein inequality for scattered data interpolation via radial basis functions, Constr. Approx. 24 (2006) 175–186.

[25] Y. Ohtake, A. Belyaev, H.P. Seidel, 3D scattered data interpolation and approximation with multilevel compactly supported RBFs, Graph. Models 67 (2005) 150–165.

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

[27] T. Sauer, Polynomial interpolation in several variables: Lattices, differences, and ideals. In: K. Jetter, M.D. Buhmann, W. Haussmann, R. Schaback, J. Stöckler (Eds.), Topics in Multivar. Approx. and Interpolation, Stud. Comput. Math. 12 (2006) pp. 191–230.

[28] C.J. Stone, Consistent nonparametric regression, Ann. Statist. 5 (1977) 595–645.

[29] C.J. Stone, Optimal global rates of convergence for nonparametric regression, Ann. Statist. 10 (1982) 1040–1053.

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

[31] G.S. Watson, Smooth regression analysis, Sankhyā Ser. A 26 (1964) 359–372.

[32] H. Wendland, Local polynomial reproduction and moving least squares approximation, IMA J. Numer. Anal. 21 (2001) 285–300.

[33] H. Wendland, Scattered data approximation, Cambridge University Press, Cambridge, UK, 2005.