ORTHOGONAL SERIES ESTIMATES ON STRONG SPATIAL MIXING DATA

JOHANNES T. N. KREBS

Abstract. We study a nonparametric regression model for sample data which is defined on an $N$-dimensional lattice structure and which is assumed to be strong spatial mixing: We use design adapted multidimensional Haar wavelets which form an orthonormal system w.r.t. the empirical measure of the sample data. For such orthonormal systems, we consider a nonparametric hard thresholding estimator. We give sufficient criteria for the consistency of this estimator. Furthermore, we derive rates of convergence for this estimator. The theorems reveal that our estimator is able to adapt to the local smoothness of the underlying regression function and the design distribution. We illustrate our results with simulated examples.

1. Introduction

We give a short review on related work and on important concepts which we shall use throughout this article. Nonparametric regression is a well established topic in statistics, new, however, is the question of how well these estimators behave when the underlying sample features a certain spatial dependence structure: we work on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ which is equipped with the random field $(X, Y) = \{(X(s), Y(s)) : s \in S\}$ which is strong spatial mixing; the definitions of these notions follow in the next section. Usually, we take $S = \mathbb{Z}^N$ for some lattice dimension $N \in \mathbb{N}$, but our discussion is not limited to that regular case; we could also allow that the random field is only partially observed at some $V \subseteq \mathbb{Z}^N$.

The random variables $X(s)$ are $\mathbb{R}^d$-valued and have equal marginal distributions denoted by the probability measure $\mu$ on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. The $Y(s)$ are $\mathbb{R}$-valued, square integrable and satisfy

$$Y(s) = m(X(s)) + \zeta(X(s)) \varepsilon(s), \text{ for each } s \in S \quad (1.1)$$

where $m, \zeta : \mathbb{R}^d \to \mathbb{R}$ are functions in $L^2(\mu)$ and the error terms $\varepsilon(s) \sim (0, 1)$ are independent of $X$ and have identical marginal distributions but may be dependent among each other such that the strong spatial mixing property remains valid. We emphasize that there is no requirement on the distribution of the error terms, e.g., a Gaussian distribution is not necessary. A possible setup is a (strictly) stationary random field $X$ and i.i.d. error terms $\varepsilon$, however, we do not require this here. Thus, we apply the classical heteroscedastic regression model to spatial data. Therefore, we give a short review on the most relevant literature in our context which treats the case for i.i.d. data observed at irregularly spaced locations: [Delouille et al. 2001] construct a soft thresholding regression estimator for univariate i.i.d. sample data. They use orthogonal design adapted Haar wavelets which in the one dimensional case even generate a multiresolution analysis (cf. Girardi and Sweldens 1997). They derive rates of convergence for H"older continuous regression functions in a model where the design variables $X$ are supposed to admit a density which has bounded support. Kohler [2003] investigates a univariate hard thresholding estimator which is constructed from a family of piecewise polynomials which are orthonormal w.r.t. the empirical design measure $\mu_n$ of data $X$ which takes values in $[0, 1]$. He derives a rate of convergence theorem under the assumption that the error terms are bounded. In a more recent article Kohler [2008] generalizes this idea to multidimensional data: he constructs an orthonormal system w.r.t. the empirical measure $\mu_n$ which consists of piecewise constant functions w.r.t. dyadic partitions of the unit cube $[0, 1]^d$. Under the assumption of sub Gaussian error terms and a bounded design distribution of $X$, he derives rates of convergence for a hard thresholding estimator.

In this article, we continue these ideas, however, we relax certain restrictions: the most important one is that the sample data is no longer required to be i.i.d. We merely assume that it is defined on a spatial lattice structure $\mathbb{Z}^N$ and that it is strong spatial mixing, i.e. the dependence between random variables vanishes if their distance on the lattice increases to infinity. The design distribution of the $X(s)$ is not restricted to a bounded domain.

Date: September 25, 2016.

2010 Mathematics Subject Classification. Primary 62G08, 62H11, 60G60.

Key words and phrases. Nonparametric nonlinear regression; empirical orthonormalization; $L^2$-consistency; rate of convergence; strong spatial mixing; spatial processes.

The author gratefully acknowledges the financial support of the Fraunhofer ITWM which is part of the Fraunhofer-Gesellschaft zur Förderung der angewandten Forschung e.V.
Furthermore, we do not require the error terms in the regression model to be sub Gaussian; we give rates of convergence for general classes of error terms for which the tail distribution is exponentially decreasing. This article is organized as follows: in the remaining section, we introduce the basic concepts. In Section 2, we present the main results of this paper: we give a general consistency theorem for our nonparametric estimator and derive a rate of convergence theorem. In Section 3, we give numerical applications and make the comparison with classical i.i.d. data. The proofs of our theorems are presented in Section 4. A, B, and C contain certain deferred proofs and further background material which proves to be useful in the broader context of random fields. We come to the main definitions:

Definition 1.1 (Random field). Let \((\Omega, \mathcal{A}, \mathbb{P})\) be a probability space, let \(V\) be an index set and let \((S_v, \mathcal{F}_v)\) be a measurable space for \(v \in V\). Let \(Z := \{Z(v) : v \in V\}\) be a set of random variables on \((\Omega, \mathcal{A}, \mathbb{P})\) such that each \(Z(v)\) takes values in \((S_v, \mathcal{F}_v)\). Then, the collection \(Z\) is called a random field.

Definition 1.2 (Homogeneous random field). Let \((\Gamma, +)\) be a group. Let \((\Omega, \mathcal{A}, \mathbb{P})\) be a probability space endowed with the random field \(Z(s) : s \in \Gamma\) where each \(Z(s)\) takes values in the same state space \((S, \mathcal{F})\). The random field is called homogeneous or stationary if for each \(n \in \mathbb{N}_+\) and for all points \(s_1, \ldots, s_n \in \Gamma\) and each translation \(t \in \Gamma\) the joint probability distribution of the collection \([Z(s_1 + t), \ldots, Z(s_n + t)]\) is identical with the joint probability distribution of \([Z(s_1), \ldots, Z(s_n)]\).

In the following we shall assume the index set \(V\) to be a subset of \(\mathbb{Z}^N\) for some positive dimension \(N \in \mathbb{N}_+\). We denote by \(\| \cdot \|_p\) the \(p\)-norm on \(\mathbb{R}^N\) and by \(d_p\) the corresponding metric for \(p \in [1, \infty]\) with the extension \(d_p(I, J) := \inf\{d_p(s, t) : s \in I, t \in J\}\) for subsets \(I, J\) of \(\mathbb{R}^N\). Furthermore, write \(s \leq t\) for \(s, t \in \mathbb{R}^N\) if and only if for each \(1 \leq k \leq N\) the single coordinates satisfy \(s_k \leq t_k\). We denote the indicator function of a set \(A\) by \(\mathbb{1}[A]\).

Definition 1.3 (Strong spatial mixing). Let \(Z(s) : s \in V\) be a random field on \((\Omega, \mathcal{A}, \mathbb{P})\) for \(V \subseteq \mathbb{Z}^N, N \in \mathbb{N}_+\). Denote for a subset \(I\) of \(V\) by \(\mathcal{F}(I) = \sigma(Z(s) : s \in I)\) the \(\sigma\)-algebra generated by the \(Z(s)\) in \(I\). Define for \(k \in \mathbb{N}_+\) the \(\alpha\)-mixing coefficient as

\[
\alpha(k) := \sup_{I,J \subseteq V, \mathcal{A} \subseteq \mathcal{F}(I), d_s(I,J) \geq k} \sup_{B \in \mathcal{B}(J)} \left| \mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B) \right|
\]

The random field is strong spatial mixing if \(\alpha(k) \to 0\) for \(k \to \infty\).

We denote by \(e_N := (1, \ldots, 1)^T\) the \(N\)-dimensional vector whose entries are equal to 1. For an \(N\)-dimensional cube in \(\mathbb{Z}^N\) that is spanned by two points \(a, b \in \mathbb{Z}^N, a \leq b\), we write \([a..b]\). The support of a function \(f : \mathbb{R}^d \to \mathbb{R}\) is \(\text{supp}(f) := \{x \in \mathbb{R}^d : f(x) \neq 0\}\). We summarize our assumptions in a regularity condition which we shall use throughout the rest of this article.

Condition 1.4 (Regularity condition for random fields). Let \(Z = \{Z(s) : s \in \mathbb{Z}^N\}\) be a random field such that each \(Z(s)\) takes values in \(\{\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d)\}\). \(Z\) is strong mixing with exponentially decreasing mixing coefficients: there are \(c_0, c_1 \in \mathbb{R}_+\), such that \(\alpha(k) \leq c_0 \exp(-c_1 k)\) for all \(k \in \mathbb{N}_+\). We shall refer to \(c_0\) and \(c_1\) as the bound on the mixing coefficients. When speaking of an increasing sequence \((n(k) : k \in \mathbb{N}_+) \subseteq \mathbb{N}_+\), we understand that \(n(k) \leq n(k+1)\) for \(k \in \mathbb{N}_+\); this sequence fulfills both

\[
\liminf_{k \to \infty} \min_{l \leq k} n_l(k) \geq e^2 \quad \text{and} \quad \liminf_{k \to \infty} \max_{l \leq k} n_l(k) = \infty \quad \text{as} \quad k \to \infty.
\]

Define the increasing index sets \(I_{n(k)} := [e_{N}..n(k)] = \{s \in \mathbb{N}_+^N : s \leq n(k)\} \subseteq \mathbb{N}_+^N\).

2. Nonlinear Hard Thresholding Regression with Wavelets

We assume the model from equation (1.1) and consider the nonlinear hard thresholding estimator for the conditional mean function \(m\): let therefore be given an increasing sequence of linear spaces \(\mathcal{F}_k\) which are dense in \(L^2(\mu)\) for functions \(g_k : \mathbb{R}^d \to \mathbb{R}\).

\[
\mathcal{F}_k = \left\{ \sum_{l=1}^{K_k} a_l g_l : a_l \in \mathbb{R} \right\}, \quad \mathcal{F}_k \subseteq \mathcal{F}_{k+1} \quad \text{and} \quad \bigcup_{k \in \mathbb{N}_+} \mathcal{F}_k \text{ is dense in } L^2(\mu). \tag{2.1}
\]

The dimension of the linear spaces which we denote by \(K^* \in \mathbb{N}_+\) depends on the index \(k \in \mathbb{N}_+\). We give an example for a standard Haar wavelet basis in \(d\)-dimensions, cf. [Benedetto 1993]: let \(d \in \mathbb{N}_+\) and let \(\varphi\) be a Haar scaling function on the real line \(\mathbb{R}\) together with the Haar mother wavelet \(\psi\). Put \(\Gamma := \mathbb{Z}^d\) and define the diagonal matrix \(M\) by \(M := 2 \text{diag}(1, \ldots, 1)\) as well as \(\xi_0 := \varphi\) and \(\xi_1 = \psi\). Set \(|M| = \text{det}(M) = 2^d\). For \(v \in [0,1]^d \setminus \emptyset\) denote the mother wavelets as pure tensors by \(\Psi_v := \xi_{v_1} \otimes \cdots \otimes \xi_{v_d}\). The scaling function is
Define the first function by
\[\Phi := \Psi_0 := \Theta^{[0]}_1 \Psi.\]
Then, \(\Phi\) and the \(\mathbb{R}\)-linear spaces \(V_j := \langle \Phi(M^j \cdot \gamma) : \gamma \in \mathbb{Z}^d \rangle\) generate an MRA of \(L^2(\mathbb{A}^d)\) and the functions \(\Psi_v, v \neq 0\), form an orthonormal basis in that
\[L^2(\mathbb{A}^d) = V_0 \oplus \bigoplus_{j \in \mathbb{Z}} W_j = \bigoplus_{j \in \mathbb{Z}} W_j \text{ where } W_j = \langle |M|^{1/2} \Psi_v(M^j \cdot \gamma) : \gamma \in \mathbb{Z}^d, v \in (0, 1]^d \setminus \{0\} \rangle.\]
Furthermore, one can show that this wavelet family is dense in \(L^2(\mu)\) where \(\mu\) is an arbitrary probability on \((\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))\). Denote by \(\Phi_{j,\gamma} := |M|^{1/2} \Phi_{j,\gamma}(M^j \cdot \gamma)\) the father wavelets w.r.t. the roughest resolution \(j_0\) and write \(\Psi_{v,j,y} := |M|^{1/2} \Psi_v(M^j \cdot \gamma)\) for the mother wavelets \(v = 1, \ldots, 2^d - 1, j \geq j_0\). Let the finest resolution index \(j_1\) be a function of \(k\), consider the \(\mathbb{R}\)-linear spaces
\[F_k := \left\{ \sum_{y \in A_{j_0}} a_{j_0,y} \Phi_{j_0,\gamma} + \sum_{j_{j_0}} \sum_{y \in A_{j_{j_0}}} b_{j_{j_0},y} \Psi_{v,j,y} : a_{j_0,y}, b_{j_{j_0},y} \in \mathbb{R} \right\}, \quad (2.2)\]
with the index sets \(A_{j,k} \subseteq \mathbb{Z}^d\). These are given as \(A_{j,k} := [-2^{-j} \cdot \gamma, 2^{-j} \cdot \gamma - 1]^d\) for a non-decreasing sequence \((\gamma_k : k \in \mathbb{N}) \subseteq \mathbb{N}_+\). If the distribution of the \((X(s))\) is bounded, it suffices to take a constant sequence such that \(U_{\gamma \in A_{j,k}} \text{supp} \Phi_{j,\gamma}\) covers the domain of the distribution. Otherwise we choose \(\gamma_k\) as increasing; we precise this in the subsequent theorems. Mark that with the definitions of the \(A_{j,k}\) we have \(U_{\gamma \in A_{j,k}} \text{supp} \Phi_{j,\gamma} = 2^{-j} \cdot [-\gamma_k, \gamma_k]^d\) for all \(j \geq j_0\).
In order to estimate the coefficients of the hard thresholding estimator, we need a set of functions which are orthonormal w.r.t. to the empirical measure \(\mu_n = |I_n|^{-1} \sum_{s \in I_n} \delta(X(s))\) for a sample \((X(s) : s \in I_n), I_n \subseteq \mathbb{Z}^d\). Delouille et al. [2001] use adaptive, orthonormal wavelets in one dimension, these are given as
\[\varphi_{j,\gamma} := \mu_n(I_{j,\gamma})^{-1/2} I_{[I_{j,\gamma}]}, \quad \psi_{j,\gamma} := \mu_n(I_{j,\gamma})^{-1/2} \left( \mu_n(I_{j+1,\gamma+1})^{1/2} \varphi_{j+1,\gamma+1} - \mu_n(I_{j+1,\gamma})^{1/2} \varphi_{j+1,\gamma} \right).\]
In the following, we outline a variant how to construct orthonormal wavelets in \(L^2(\mu)\) in higher dimensions; however, these fulfill no longer the usual scaling equations which are satisfied in the case of the Lebesgue measure because the empirical measure \(\mu_n\) on \(\mathbb{R}^d\) is not a product measure if \(d > 1\) and a partition of \(\mathbb{R}^d\) into Cartesian products of intervals does in general not satisfy that each partition element contains the same number of sample points. We use the Gram-Schmidt rule w.r.t. the empirical scalar product
\[\langle f, g \rangle_n := \frac{1}{|I_n|} \sum_{s \in I_n} f(X(s)) g(X(s)) = \int_{\mathbb{R}^d} f \ g \ d\mu_n.\]
As the father wavelets partition \(\mathbb{R}^d\), we use the following ansatz to construct these balanced wavelets: let there be given a cube \(A = A_1 \times \ldots \times A_d\) where the \(A_i = [a_i, b_i] \subseteq \mathbb{R}\) are finite intervals. Let the family of standard Haar wavelets \(\{\psi_k : k \in [0, 1]^d\}\) from above be translated and rescaled such that \(\Phi = \Psi_0 = \mathbb{I}[A]\). Let there be given a linear ordering of the \(2^d\) basis functions as follows
\[(0, \ldots, 0) \mapsto 0, \quad (1, 0, \ldots, 0) \mapsto 1, \quad (0, 1, 0, \ldots, 0) \mapsto 2, \ldots, (1, \ldots, 1) \mapsto 2^d - 1.\]
Define the first function by \(f_0 := \mu_n(A)^{-1/2} \mathbb{I}[A]\). Denote the left half of an interval \(A\) by \(A_L := [a_i, (a_i + b_i)/2]\) and the right one by \(A_R := [(a_i + b_i)/2, b_i]\). Furthermore set \(A_{-i} := \prod_{j \neq i} A_j\). We call a mother wavelet balanced if it integrates to zero w.r.t. the empirical measure \(\mu_n\). Define the second function which is the first balanced and orthonormal wavelet as
\[f_i := \frac{\mu_n(A^k \times A_{-i})}{\mu_n(A^k \times A_{-i})} - \frac{\mu_n(A^k \times A_{-i})}{\mu_n(A^k \times A_{-i})} \frac{1}{\mu_n(A^k \times A_{-i})} \left[ A^k \times A_{-i} \right].\]
Assume that w.r.t. \(\mu_n\) the functions \(f_0, \ldots, f_{d-1}\) are orthonormal and that \(f_{d-1}, \ldots, f_{d-1}\) are additionally balanced for \(1 \leq u \leq 2^d - 1\). Note that some of these \(f_u\) might be zero in \(L^2(\mu_n)\). Then the function \(f_{d+1}\) which one obtains from the Gram-Schmidt rule
\[f_{d+1} := \Psi_{d+1} - (\Psi_{d+1}, f_u) f_u - \ldots - (\Psi_{d+1}, f_1) f_1 - (\Psi_{d+1}, f_0) f_0\]
is balanced (and orthogonal to \((f_0, \ldots, f_{d-1})\)). If \(f_{d+1}\) is not zero w.r.t. \(\mu_n\), normalize this function and obtain \(f_{d+1}\). Repeating this step until \(u = 2^d - 1\) and obtain \(\tilde{K} \leq 2^d - 1\) balanced and orthonormal wavelets \(f_1, \ldots, f_{\tilde{K}}\). Note that each of these functions is constant on a subcube of the kind \(A_u = \prod_{j=1}^{d} A_{u_j}\) with \(u_j \in [L, R]\). Hence, executing the same procedure on \(A_u\), one obtains a new set of \(\tilde{K}_{\tilde{K}} \leq 2^d - 1\) orthonormal and balanced wavelets,
call them \(g_1,\ldots,g_{K^*}\) which are orthogonal to \(f_{j_1},\ldots,f_{j_k}\). Consequently, if one repeats this procedure ad infinitum one can construct a wavelet family on \(A\) (and its dyadic subcubes) which is orthonormal. In particular, given a partition of \(\mathbb{R}^d\) by cubes of the kind of \(A\) which have finite Lebesgue measure and an empirical measure \(\mu_n\) one can construct such a wavelet family on entire \(\mathbb{R}^d\), \(\{f_{j_0,j_1,\ldots,j_{d-1}}, j_0, j_1, \ldots, j_{d-1}, \gamma \in \mathbb{Z}^d\}\). Furthermore, for the function spaces as \(2.2\), we have

\[
\mathcal{F}_k = \{f_{j_0,j_1,\ldots,j_{d-1}} : v = 1,\ldots,2^d - 1, j = j_0,\ldots,j_1 = 1, \gamma \in \mathbb{Z}^d\} \text{ in } L^2(\mu_n) \text{ a.s.}
\]

where the space is spanned by at most \(K^*(k) = (2 \cdot 2^{(k-j_{d-1})})^{d}\) wavelets. Kohler constructs an alternative ONB in [Kohler 2008], which still has the property that the functions are balanced w.r.t. \(\mu_n\), however, each function vanishes on a larger set than our corresponding function.

In the following, we use the generalized notion that the \(\mathcal{F}_k\) are given as the linear span of \(K^*(k)\) deterministic functions as in \(2.1\) and that there are random orthonormal functions \(f_1,\ldots,f_k \in \mathcal{F}_k\) such that \(\mathcal{F}_k = \langle f_1,\ldots,f_k \rangle \) in \(L^2(\mu_n)\) a.s. Furthermore, these (random) functions satisfy for each \(1 \leq u \leq K^*\) the relation

\[
\langle f_1,\ldots,f_u \rangle = \langle g_1,\ldots,g_u \rangle \text{ in } L^2(\mu_n).
\]

Note that some functions might be zero w.r.t. the empirical measure, so \(K \leq K^*\) a.s. For such an orthonormal system we estimate the coefficients consistently: set \(J := \{1,\ldots,\hat{K}\}\) and \(B = \{f_j(X(s)) \mid s \in I_{j_1,\ldots,j_{d-1}}\}\), by construction \(|I_{j_1,\ldots,j_{d-1}}|^{-1} B^T B = I\) is the identity matrix. Then define the estimates of the coefficients of the regression function as

\[
\frac{1}{|I_n|} B^T B a^\ast = \frac{1}{|I_n|} B^T Y \Leftrightarrow a^\ast = \frac{1}{|I_n|} \sum_{j \in I_{n}} f_j(X(s)) Y(s).
\]

Let \(\{\lambda_k : k \in \mathbb{N}\} \subseteq \mathbb{R}_+\) be the hard thresholding sequence which converges to zero: define the nonlinear thresholded estimator \(m_{k,J}\) with the notions

\[
J' := \{j \in J : |a^\ast_j| > \lambda_k\} \quad \text{and} \quad m_{k,J} := \sum_{j \in J'} a^\ast_j f_j \text{ for } \hat{J} \subseteq J.
\]

Then, by elementary reasoning, cf. [Kohler 2003], hard thresholding corresponds to \(L^0\)-penalized least squares which means that for a subset \(\hat{J}\) of \(J\) the penalizing term \(\text{pen}_k(\hat{J}) := \lambda_k^2 |\hat{J}|\) satisfies the relation

\[
\frac{1}{|I_n|} \sum_{j \in I_{n}} |m_{k,J}(X_s) - Y_s|^2 + \text{pen}_k(J') = \min_{\hat{J} \in J} \left\{ \frac{1}{|I_n|} \sum_{j \in I_{n}} |m_{k,J}(X_s) - Y_s|^2 + \text{pen}_k(\hat{J}) \right\}.
\]

We write for short \(m_k := m_{k,J}\) for the minimizing function. Define for \(L > 0\) the truncation operator as \(T_L(x) := \max(\min(y,L),-L)\). Let \(\beta_k : k \in \mathbb{N}_+\) be a real-valued, non decreasing truncation sequence which converges to infinity. In order to render the estimator robust against deviations in the data \(\{X(s) : s \in I_n\}\), we consider the truncated hard thresholding least-squares estimator

\[
\hat{m}_k := T_{\beta_k} m_k.
\]

Furthermore, for a function \(f \in \mathcal{F}_k\), which has a unique representation w.r.t. the functions \(\{f_1,\ldots,f_k\}\) in \(L^2(\mu_n)\), we write \(\text{pen}_k(f)\) for the number of nonzero coefficients in this representation multiplied by \(\lambda_k^2\). Then, the penalizing term \(\text{pen}_k(f)\) is stochastic and bounded by \(K^*(k) \lambda_k^2\). \(\hat{m}_k\) is a consistent estimator under the conditions in the next theorem:

**Theorem 2.1** (Consistency hard thresholding). Let Condition \(2.1(a)\) and \(2.2(b)\) be satisfied for \(L = \mathbb{R}^N\) and for a random field \(\{X(s), Y(s) : s \in \mathbb{Z}^N\}\) which fulfills the regression problem from \(1.1\) and where the \(Y(s)\) are square integrable. Let the function classes be given as linear spaces which are as in \(2.2\). In particular, if the spaces are given by \(2.2\) and \(2.3\) and if the distribution of the \(X(s)\) is unbounded, let \(\lim_{k \to \infty} \beta_k = \infty\). The estimator \(\hat{m}_k\) is weakly universally consistent in that

\[
\lim_{k \to \infty} \mathbb{E} \left[ \int_{\mathbb{R}^d} |\hat{m}_k - m|^2 \, d\mu_X \right] = 0
\]

if both

\[
K^*(k) \lambda_k^2 \to 0 \quad \text{and} \quad \beta_k^2 K^*(k) \log \beta_k \prod_{i=1}^{N} \log n_i(k) \left( \prod_{i=1}^{N} n_i(k) \right)^{1/(N+1)} \to 0 \text{ as } k \to \infty.
\]
If the function spaces are given by (2.2) and (2.3), $K^*(k) = (2^{h(k)})^{-1} w_k^d$. Furthermore, the nonlinear wavelet estimator $\hat{m}_k$ is strongly universally consistent in that

$$\lim_{k \to \infty} \int_{\mathbb{R}^d} |\hat{m}_k - m|^2 \, d\mu_X = 0 \quad a.s.$$ if in addition $(Y(s) : s \in \mathbb{Z}^N)$ is stationary and if for some positive $\delta > 0$

$$\beta_k^2 (\log k)^{1+\delta} \left( \prod_{i=1}^N \log n_i(k) \right)^{1/(N+1)} \to 0 \text{ as } k \to \infty.$$ Under the condition that the error terms are exponentially decreasing, we can derive a rate of convergence theorem. Therefore, we introduce a piece of notation: for the constructed orthonormal wavelet system, we define sets of partitions of $D_k := \bigcup_{y \in A_{n,k}} \text{supp} \Phi_{h,y}$ which we denote by $\Pi_u$ for $1 \leq u \leq u_{\text{max}}$ where

$$u_{\text{max}} := 1 + (2w_{k})^d [(2^{d(j-h)} - 1)/(2^d - 1)] = 1 + (2w_{k})^d \sum_{j=0}^{h_j-1} 2^d.$$

Set $\Pi_1 := \{\text{supp} \Phi_{h,y} : y \in A_{n,k}\}$. Let $\Pi_1, \ldots, \Pi_n$ be constructed. Then $\pi \in \Pi_{\text{max}}$ and only if there is a partition $\pi'$ of $D_k$ and a dyadic cube $A$ such that

$$\pi = \pi' \setminus \{A \cup \{A_{i_1}^n \times \ldots \times A_{i_u}^n : u_i \in [L,R]\}$$

and each element of the partition $B = B_1 \times \ldots \times B_n$ consists of intervals of a length greater or equal than $2^{-h}$. This means $\Pi_{u+1}$ consists of all partitions of $\Pi_u$ which are refined by partitioning exactly one element into $2^d$ equidistance cubes and each partition element has a diameter w.r.t. the $\infty$-norm of at most $2^{-h}$. Note that $u_{\text{max}}$ is the maximal index such that $\Pi_{u+1}$ contains finer partitions than $\Pi_u (u < u_{\text{max}})$ and that $\Pi_{u_{\text{max}}}$ contains a single partition, namely, $\Pi_{u_{\text{max}}} = \{\text{supp} \Phi_{h,y} : y \in A_{n,k}\}$.

Hence, starting out from $\Pi_1$, there are exactly $(2w_k)^d[(2^{d(j-h)} - 1)/(2^d - 1)]$ possible refinements; we have to add a 1 to this number for the index $u_{\text{max}}$ as we start counting at 1 and not at 0. We denote the functions which are constant w.r.t. a partition $\pi \in \Pi_u$ for some $1 \leq u \leq u_{\text{max}}$ by $F_c \circ \pi$.

**Theorem 2.2** (Rate of convergence in case of exponentially decreasing errors). Let Condition 1.2 (a) and (b) be satisfied for the random field $(X,Y)$ and the index set $I = \mathbb{Z}^N$. Furthermore, let the regression function $m$ and the conditional variance function $\varsigma^2$ be essentially bounded. Put $B := \|m\|_{\infty}$ and $\beta_k := B$ for all $k \in \mathbb{N}_+$.

The error terms $e(s)$ fulfill the exponential tail condition

$$\mathbb{P}(|e(s)| > z) \leq \kappa_0 \exp (-\kappa_1 z^\tau), \text{ for some } \kappa_0, \kappa_1, \tau \geq 0.$$

The thresholding sequence $(\lambda_k : k \in \mathbb{N}_+)$ and the growth of the basis functions satisfy the relations

$$K^*(k)^2 \lambda_k^2 \to 0 \text{ and } (K^*(k))^{1+2/\tau} \left( \prod_{i=1}^N \log n_i(k) \right)^{3+4/\tau} \left| m_k \right|^{1/(N+1)} \to 0 \text{ as } k \to \infty.$$ Then there is a constant $C \in \mathbb{R}_+$ which only depends on $B$, $\|\varsigma\|_{\infty}$, the lattice dimension $N$, the bound on the mixing coefficients and the parameters of the tail distribution of the error terms such that the $L^2$-error satisfies for the general sequence of function spaces $\mathcal{F}_k$ from (2.1) for all $k \in \mathbb{N}_+$ the relation

$$\mathbb{E} \left[ \int_{\mathbb{R}^d} |\hat{m}_k - m|^2 \, d\mu_X \right] \leq 4 \min_{1 \leq \lambda \leq 2^{K^*(k)}} \left\{ \lambda_k^2 \lambda + \inf_{f \in \mathcal{F}_k} \int_{\mathbb{R}^d} |f - m|^2 \, d\mu_X \right\} \left( \prod_{i=1}^N \log n_i(k) \right)^{3+4/\tau} \left| m_k \right|^{1/(N+1)} + C (K^*(k))^{1+2/\tau} \left( \prod_{i=1}^N \log n_i(k) \right)^{3+4/\tau} \left| m_k \right|^{1/(N+1)}.$$

In the particular case of the constructed orthonormal wavelet system from (2.3), this bound can be refined as follows

$$\mathbb{E} \left[ \int_{\mathbb{R}^d} |\hat{m}_k - m|^2 \, d\mu_X \right] \leq 4 \min_{1 \leq \lambda \leq 2^{K^*(k)}} \left\{ \lambda_k^2 (2^d - 1)(\mu - 1) + (2w_{k})^d \right\} + \min_{\pi \in \Pi_u, f \in \mathcal{F}_k} \int_{\mathbb{R}^d} |f - m|^2 \, d\mu_X \right.$$
Under the more severe restriction of a bounded regression function $m$ and bounded error terms, the rate of convergence of $\hat{m}_k$ can be improved. We only state the result for the general function spaces from (2.1). The application to the orthonormal wavelet system from (2.2) and (2.3) is straightforward.

**Theorem 2.3** (Rate of convergence in case of bounded errors). **Under the conditions of Theorem 2.2**. Additionally, let the error terms $s(s)$ be essentially bounded by $L \in \mathbb{R}^+$, i.e., $\|s(s)\| \leq L$ for all $s \in \mathbb{R}^N$. Then there is a constant $C \in \mathbb{R}_+$ which only depends on $B, L, \|\xi\|_i$, the lattice dimension $N$ and the bound on the mixing coefficients such that the estimator $\hat{m}_k$ truncated at $\pm (B + \|\xi\|_i L)$ satisfies for all $k \in \mathbb{N}_+$

\[
\mathbb{E} \left[ \int_{\mathbb{R}^d} \left( \|\hat{m}_k - m\|^2 \cdot d\mu_x \right) \right] \leq 4 \min_{1 \leq \ell \leq k} \left\{ \lambda_\ell^2 + \inf_{f \in S_{\ell}^k} \int_{\mathbb{R}^d} |f - m|^2 \cdot d\mu_x \right\} + C \cdot K(k) \left( \prod_{i=1}^{\ell} \log n_i(k) \right)^3 \left\| \hat{m}_k \right\|_1^{1/(N+1)}.
\]

We give two examples of application for the isotropic Haar basis and an $(A, r)$-Hölder continuous regression function $m$.

**Corollary 2.4** (Hölder continuous functions). **Under the conditions of Theorem 2.2 and the orthonormal wavelet system from equation (2.3)**. Let $m$ be $(A, r)$-Hölder continuous. Furthermore, assume that the Euclidean norm of $X$, $\|X\|_{\mathbb{R}^N}$, is integrable w.r.t. the probability measure $\mathbb{P}$ for some $q \in \mathbb{R}_+$. Let $C_0, C_1, C_2 \in \mathbb{R}_+$ be constants. Define the parameters as

\[
\bar{R}(n) = \left( \prod_{i=1}^{N} \log n_i \right)^{3+\frac{4}{r}} \left\| \mathcal{H} \right\|_{1/(N+1)},
\]

\[
w_k = C_0 \exp \left\{ \frac{2r}{2rd(1 + 2/r) + q(2r + d(1 + 2/r))} \log \bar{R}(n(k)) \right\},
\]

\[
\lambda_k^2 = C_1 \exp \left\{ \frac{2r + d}{2rd(1 + 2/r) + q(2r + d(1 + 2/r))} \log \bar{R}(n(k)) \right\}
\]

\[
\text{and } j_1(k) = \left\lfloor C_2 - \frac{q}{2rd(1 + 2/r) + q(2r + d(1 + 2/r))} \log \bar{R}(n(k)) \right\rfloor.
\]

Then the rate of convergence is at least

\[
\mathbb{E} \left[ \int_{\mathbb{R}^d} \left( \|\hat{m}_n - m\|^2 \cdot d\mu_x \right) \right] \in O \left( \bar{R}(n(k))^{2q/[2rd(1+2/r) + q(2r + d(1+2/r))]} \right).
\]

If the distribution of $X$ is bounded and if

\[
\lambda_k^2 = C_1 \exp \left\{ \frac{2r + d}{2rd(1 + 2/r) + q(2r + d(1 + 2/r))} \log \bar{R}(n(k)) \right\}
\]

\[
\text{and } j_1(k) = \left\lfloor C_2 - \frac{1}{2rd(1 + 2/r) + q(2r + d(1 + 2/r))} \log \bar{R}(n(k)) \right\rfloor.
\]

the estimator achieves a rate of at least $O \left( \bar{R}(n(k))^{2r/[2r + d(1 + 2/r)]} \right)$.

**Proof.** It remains to compute the approximation error, we choose $j_0 = 0$ as the roughest resolution: there is a function $f \in \mathcal{F}_k$ (from equations (2.2) and (2.3)) which is piecewise constant on dyadic $d$-dimensional cubes of edge length $2^{j_k}$ with values

\[
f(x) = m(x^*(\gamma)) \quad \text{for } x \in \left( \frac{\gamma_1, \ldots, \gamma_d}{2^{j_k}}, \frac{(\gamma_1, \ldots, \gamma_d) + e_N}{2^{j_k}} \right)
\]

\[
\text{and } x^*(\gamma) \in \left( \frac{\gamma_1, \ldots, \gamma_d}{2^{j_k}}, \frac{(\gamma_1, \ldots, \gamma_d) + e_N}{2^{j_k}} \right) \cap \text{dom}(m)
\]

where $\gamma_i \in \mathbb{Z}$ for $i = 1, \ldots, d$. 

\[\]
In case of an unbounded distribution of $X$, the domain of $f$ is the cube $[-w_k, w_k]^d$ and in case of a bounded distribution, $f$ is supposed to be defined on the entire domain of $X$. The approximation error is at most
\[
\int_B |f - m|^2 \ d\mu_X \leq \sup_{\text{dom}(f)} |f - m|^2 + ||m||_\infty^2 \ P(||X||_\infty \geq w_k) \leq A^2 \ 2^{-j_0(k)\tau} + ||m||_\infty^2 \ E[||X||_\infty] \ / w_k^{2\tau}.
\]

If the distribution is bounded the second term on the RHS in the last expression is zero. The growth rates of $j_0$, $\lambda$, $w_k$ equalize the asymptotic rates of the error terms in both cases.

Corollary 2.4 reveals that the lattice dimension $N$ and the data dimension $d$ influence the rate of convergence negatively whereas the parameters $\tau$ and $r$ have a positive influence. The reason for the negative impact of $N$ is that the dependence in our model can spread in every dimension of the lattice, hence, observing data on an additional lattice dimension can become more and more redundant. The effect of the data dimension $d$ is the well-known curse of dimensionality. The positive impact of the parameter $\tau$ is clear because an increase in $\tau$ goes hand in hand with a reduced uncertainty. Similarly, an increase in $r$ means a smoother regression function which can be better approximated by finite linear combinations.

The results from Corollary 2.4 remind on the classical rate for $(\lambda, \lambda)$-Hölder continuous regression functions defined on a bounded domain and an i.i.d. sample of $n$ observations where the $L^2$-risk decreases essentially at a rate of $n^{-2/(2r+d)}$ times a logarithmic factor. Kohler [2008] considers a multivariate set-up for an i.i.d. sample with $\{X_1, \ldots, X_n\} \subseteq [0, 1]^d$, a bounded regression function $m$ and sub-Gaussian error terms. This corresponds to our scenario of bounded $X(s)$ and a decay rate for the error terms $\gamma \geq 2$. For the hard thresholding sequence $\lambda_n = C \sqrt{\log n / n}$ he obtains a rate of $(\log n / n)^{2/(2r+d)}$. In comparison, our rate shows two additional correction factors which come from the dependence relations: instead of $d$ we have $d(1 + \tau) / \tau$ which is larger and additionally the exponent is multiplied by the standard correction factor which depends on the lattice dimension $N$. Especially for a bounded distribution, if $\tau \geq 2$ and if $N = 2$ and for the canonical sequence $n(k) := k \epsilon n$, we achieve a rate of $(\log k)^3 / (k^2 \epsilon^2)^{2/(2r+2d)}$ for a sample of size $k^2$. Kohler [2003] investigates rates of convergence for a hard thresholding estimator which is constructed from piecewise polynomials for i.i.d. sample data where the distribution of $X$ is in the unit interval and $Y$ is bounded by $L \in \mathbb{N}$. In this setting for an $(\lambda, \lambda)$-Hölder continuous function the error decreases essentially at a rate of $(\log n)^3 / (n^2)^{2/(2r+1)}$ for an i.i.d. sample $\{X_1, Y_1, \ldots, X_n, Y_n\} \subseteq [0, 1] \times [-L, L']$. This corresponds again to the classical rate. Compare this for instance, to our case of an underlying mixing sample on a one dimensional chain and a one dimensional and bounded $X$ as well as bounded $Y$ (which corresponds formally to $\tau = \infty$): for observations $X_1, \ldots, X_n$ the rate of convergence is at least $(\log n)^3 / n^{2/(2r+1)}$ for a sample of size $n$.

[Delouille et al. 2001] investigate the soft thresholding estimator for adaptive wavelets. They require the existence of a compactly supported one dimensional density for the distribution of $X$ and the existence of all moments of the error terms; the latter is less restrictive than our exponentially decreasing tail condition. For an adaptive soft thresholding estimator and an i.i.d. sample, they investigate a similarly defined rate of convergence (w.r.t. the empirical distribution) which is in $O(\log n / n)^{2/(2r+1)}$.

Next, consider piecewise $(\lambda, \lambda)$-Hölder continuous regression functions: there is a finite partition $\cup_{i=1}^S U_i \ (S < \infty)$ of the domain of $m$ such that $m$ is $(\lambda, \lambda)$-Hölder continuous on each $U_i$.

**Corollary 2.5** (Piecewise smooth functions). Under the same conditions as Corollary 2.4 Additionally, assume that $X$ takes values in a bounded domain $D = [-w, w]^d \ (w \in \mathbb{N}_+)$. The regression function $m$ is bounded by $B$ and is piecewise $(\lambda, \lambda)$-Hölder continuous such that for all $j \geq 0$ the condition
\[
\#\{y \in [-2^j w, 2^j w - 1] : m(\lambda, \lambda) \} \leq C 2^{d(d-1)/2}
\]

is satisfied for some constant $C$. The distribution of the $X(s)$ admits a density $g$ which is essentially bounded. Define for some $C_0, C_1 \in \mathbb{R}_+$ the thresholding sequence and the resolution by
\[
\lambda_j = C_0 \exp \left\{ \frac{1 \wedge 2^j + d}{1 \wedge 2^j + d(1 + 2/\tau)} \log R(n(k)) \right\},
\]
\[
j_1 = \left[ C_1 - \frac{1}{1 \wedge 2^j + d(1 + 2/\tau)} \log 2 \log R(n(k)) \right].
\]

Then, the $L^2$-error is in $O \left( R(n(k))^{1 + 2r} / [1 \wedge 2^j + d(1 + 2/\tau)] \right)$. 

Proof: The proof is similar: let there be given a resolution \( j_1 \) and fix the roughest resolution as \( j_0 := 0 \) which correspond to a partition \( \pi \) of the cube \([-w, w]^d\). Define \( f \) as in the proof of Corollary 2.4. Denote by \( D_{\alpha_1(j_1)} \) the set of dyadic cubes of edge length \( 2^{-j_1} \) which contain points where \( m \) is not continuous. Then the approximation error for a resolution up to \( j_1 \) is at most

\[
\int |f - m|^2 \, d\mu_X \leq A^2 2^{-2j_1r} + \int_{D_{\alpha_1(j_1)}} |f - m|^2 \, d\mu_X \leq A^2 2^{-2j_1r} + (2B)^2 \|g\|_{\infty} 2^{-j_1d} C 2^{d(d-1)h},
\]

here we use the regularity condition on the discontinuities from (2.6). The definitions of \( j_1 \) and \( \lambda_k \) equalize the individual error terms.

Before we discuss this result, consider the requirement in equation (2.6): if \( d \geq 2 \), we consider the boundary \( \partial U \) of one such partitioning element \( U \in \{U_i : i = 1, \ldots, S\} \). Let \( \partial U \) be a finite union of smooth hypersurfaces, \( \partial U = /\bigcup_{j=1}^r H_r \), where each \( H_r \) can be represented as the graph of a \( C^1 \)-function: pick one such hypersurface \( H \) which has w.l.o.g. the following representation and location in \( \mathbb{R}^d \)

\[
H = \{(x-d, h(x-d)) : x-d \in B \} \subseteq [0, 1]^d,
\]

and \( h : B \to \mathbb{R} \) such that \( \nabla h \) can be extended to a continuous function on \( \overline{B} \). Let there be given the dyadic partition \( \pi_j \) of the unit cube \([0, 1]^d\) in \( 2^j \)-equivolume dyadic subcubes of edge length \( 2^{-j} \). Consider a partition element \( \Box \in \pi_j \) which lies in the plane where the \( d \)-th dimension is zero and intersects with \( B \), i.e., \( \Box \cap B \neq \emptyset \). Then the number of partition elements in \( \pi_j \) which intersect with the image of \( \Box \cap B \) under \( h \) is bounded: indeed, use approach "steepest ascent times longest path" which yields a maximal "height". Divide this number by the edge length of the cubes, this yields the approximate number of these partitioning elements. More formally and more precisely,

\[
\left| \{ \Box \in \pi_j : \Box \cap (\Box \cap B \times h(\Box \cap B)) \neq \emptyset \} \right| \\
\leq \left( 2^{-j} \sqrt{d-1} \cdot \max_{x \in \Box} \|\nabla h(x)\|_2 \right) 2^{-j} + 1 \leq C
\]

for a constant \( C \) which is independent of \( j \in \mathbb{N}_+ \). Hence, the total number of partition elements \( \Box \in \pi_j \) which intersect with \( H \) is in \( O(2^{d(d-1)j}) \). Consequently, the total number of partition elements which intersect with \( \partial U \) is in \( O(2^{d(d-1)j}) \) as required in (2.6).

In light of this interpretation of the condition in (2.6), Corollary 2.5 illustrates that given there are discontinuities, an increase in the smoothness increases the rate of convergence only as long as \( r < 1/2 \), otherwise, if \( r \geq 1/2 \), the negative impact at the borders \( \partial U \), is too prominent and dominates the approximating property on the parts of \( D \) where \( m \) is smooth.

3. Applications to simulated data

In Example 3.6 we discuss an algorithm which enables us to simulate the multivariate normal distribution as realizations of a Markov random field on a finite graph. We continue with this idea: let \( G = (V, E) \) be a finite graph with nodes \( v_1, \ldots, v_N \). We can simulate a \( d \)-dimensional Markov random field \( Z \) on \( G \) where the marginals of the single components \( Z(v) : v \in V \) are standard normally distributed and dependent among each other according to the structure of the graph. We choose a lattice in two dimensions: the edge length is \( N = 70 \) such that we have 4900 observations in total. For the simulation of the field \( Z \) we run a Markov chain of \( M_1 = 15k \) iterations, cf. Appendix C. Furthermore, certain components \( Z_i \) can be simulated as dependent with copulas: we simulate three standard normally distributed random fields \( Z_1, Z_2 \) and \( Z_3 \) where \( Z_1 \) is independent and \( Z_1 \) and \( Z_2 \) have a correlation of approximately 0.45. The parameter \( \eta \) which describes the dependence within a random field \( Z_i \) is chosen as \( \eta = [0.2, -0.23, 0.1] \). Note that \( |\eta_1| \approx 0.23 \) constitutes a strong dependence whereas \( \eta_1 \approx 0 \) indicates independence. A positive (resp. negative) \( \eta_1 \) means a positive (resp. negative) correlation among random variables which are neighbors in the graph. In this case the admissible range for \( \eta_1 \) is very close to \([-0.25, 0.25] \) which is the parameter space for a lattice wrapped on a torus. In the next step, we construct from \( Z \) a two dimensional random field \( [X_1(v), X_2(v)) : v \in V \) and a field with error terms \( \varepsilon(v) : v \in V \) as follows: for the error terms, we choose the independent component \( Z_3 \), thus, these are standard normally distributed. For the field \( (X_1, X_2) \) we retransform \( (Z_1, Z_2) \) as follows: (a) We retransform each \( Z_i \) with a the inverse distribution function of the standard normal distribution to the interval \([-1,1] \) and
obtain $X_i$, hence, there remains a correlation between $X_1$ and $X_2$. (b) We retransform $Z_i$ as in (a), additionally, we transform linearly all $X_2$ which are less than 0.1 onto $[0,0.5]$ and the remaining $X_2$ onto $[0.5,1]$, i.e.

$$X_2 \sim \begin{cases} 0.5 \quad &\text{if } X_2 < 0.1 \\ 0.1 \cdot X_2 &\text{if } 0.1 \leq X_2 \leq 0.5 \\ 0.5 - 0.1 \cdot X_2 &\text{if } X_2 > 0.5 \end{cases}$$

$$X_2 \sim \begin{cases} 0.5 \quad &\text{if } X_2 < 0.1 \\ 1 - 0.1 \cdot (X_2 - 0.5) &\text{if } 0.5 \leq X_2 \leq 1 \\ 0.5 - 0.1 \cdot (X_2 - 0.5) &\text{if } X_2 > 1 \end{cases}.$$

Hence, in (a) the $(X_1, X_2)$ are approximately uniformly distributed on $[-1,1]^2$ with a correlation of 0.43. In (b) the lower half of $[-1,1]^2$ contains approximately only 10% of the data and the upper half 90%.

The regression functions are given as

$$m_1(x) := 4 + 6x_1^2 - 4x_2^2 \quad \text{and} \quad m_2(x) = m_1(x) \| \|x\|_2 \leq 0.5 - m_1(x) \| \|x\|_2 > 0.5.$$

All in all, we consider four different set-ups of the kind $Y(v) = m(X(v)) + \epsilon(s)$. For the estimation procedure, we choose our Haar basis from (22) and (23).

Let there be given a simulated sample $(X(v), Y(v) : v \in V)$. We partition the data in a learning and in a verification sample: $(X(v), Y(v) : v \in V_L)$ and $(X(v), Y(v) : v \in V_V)$, where we require that at least $V_L$ is a convex set. Here the learning sample comprises approximately 80% of the data which corresponds to the approach of Kohler [2008]. Then we estimate the regression function from the learning sample for a given truncation $L$ and a threshold $\lambda$ and obtain the estimator $\hat{m}_L$. We compute the approximate $L^2$-error with Monte Carlo integration over the verification sample, i.e. $L^2(\hat{m}_L) \approx |V_V|^{-1} \sum_{v \in V_V} |\hat{m}_L(X(v)) - m(X(v))|^2$.

| $\lambda$ | Estimates on 2 dimensional lattice | Independent reference estimates |
|----------|----------------------------------|---------------------------------|
|         | (a) w/ $m_1$ | (b) w/ $m_1$ | (a) w/ $m_2$ | (b) w/ $m_2$ | (a) w/ $m_1$ | (b) w/ $m_1$ | (a) w/ $m_2$ | (b) w/ $m_2$ |
| 0.0     | 0.372 (0.031) | 0.278 (0.027) | 1.289 (0.253) | 1.151 (0.233) | 0.368 (0.029) | 0.277 (0.028) | 1.286 (0.255) | 1.138 (0.236) |
| 0.02    | 0.315 (0.030) | 0.246 (0.027) | 1.243 (0.254) | 1.126 (0.233) | 0.314 (0.027) | 0.243 (0.026) | 1.241 (0.256) | 1.112 (0.236) |
| 0.04    | 0.218 (0.022) | 0.192 (0.020) | 1.176 (0.255) | 1.092 (0.235) | 0.217 (0.019) | 0.192 (0.021) | 1.174 (0.259) | 1.077 (0.239) |
| 0.06    | 0.233 (0.019) | 0.215 (0.019) | 1.226 (0.257) | 1.138 (0.238) | 0.232 (0.018) | 0.213 (0.018) | 1.224 (0.258) | 1.121 (0.243) |
| 0.08    | 0.295 (0.024) | 0.261 (0.023) | 1.340 (0.262) | 1.226 (0.247) | 0.295 (0.024) | 0.260 (0.022) | 1.341 (0.262) | 1.211 (0.250) |
| 0.10    | 0.375 (0.029) | 0.317 (0.029) | 1.475 (0.266) | 1.342 (0.255) | 0.374 (0.028) | 0.315 (0.027) | 1.477 (0.263) | 1.328 (0.258) |

Table 1. $L^2$-error of regression problems 1 - 4: the estimated mean and in brackets the estimated standard deviation for a resolution $j = 5$. In all cases the mean of the estimated $L^2$-error is minimized for the hard thresholding value $\lambda = 0.04$.

We repeat this step $M_2 = 1000$ times and obtain for a given threshold $\lambda$ an approximate mean and standard deviation of the $L^2$-error. Then we choose the threshold $\lambda$ which minimizes the $L^2$-error in the mean. We
give our results in Table 1 there we additionally give the results for independent reference samples. For the independent samples the design distribution of $X$ has in both cases correlations which match those of the respective dependent samples. Note that in all cases the hard thresholding value $\lambda = 0.04$ yields the best fit. Furthermore, the $L^2$-error measure for independent samples is nearly always better than for the corresponding dependent samples. Figures 1 and 2 depict the best fit in each case for the dependent sample: one finds that the regression estimator is able to adapt both to the local smoothness of the underlying regression function and to the design distribution.
Figure 1. True regression function $m_1$ (top) and estimates $\hat{m}_k$ with uniform data $X$ (middle) and nonuniform data (bottom). Note that the partition which is chosen depends on the data and on the local smoothness of the function.
Figure 2. True regression function $m_2$ (top) and estimates $\hat{m}_k$ with uniform data $X$ (middle) and nonuniform data (bottom). Again, the partition depends on the data and on the local smoothness of the function.
4. Proofs

We write $A, A_i, \tilde{A}_i$, resp. $C, C_i$ and $\tilde{C}_i$ for constants whose values are not necessarily the same.

Proof of Theorem 2.1. We have with the defining property of $m$ and the properties of the conditional expectation for an independent observation $(X'(e_N), Y'(e_N))$

$$\mathbb{E} \left[ |\hat{m}_k(X'(e_N)) - Y'(e_N)|^2 \mid X(I_{n,k}), Y(I_{n,k}) \right] = \mathbb{E} \left[ |\hat{m}_k(X'(e_N)) - m(X'(e_N))|^2 \mid X(I_{n,k}), Y(I_{n,k}) \right] + \mathbb{E} \left[ |m(X'(e_N)) - Y'(e_N)|^2 \right].$$

Thus,

$$\int_{\mathbb{R}^d} |\hat{m}_k - m|^2 \, d\mu_X = \mathbb{E} \left[ |\hat{m}_k(X'(e_N)) - Y'(e_N)|^2 \mid X(I_{n,k}), Y(I_{n,k}) \right] - \mathbb{E} \left[ |m(X'(e_N)) - Y'(e_N)|^2 \right].$$

Since $\mathbb{E} \left[ |m(X'(e_N)) - Y'(e_N)|^2 \right]$ is constant for all $k$ and $a - b = (\sqrt{a} - \sqrt{b})(\sqrt{a} + \sqrt{b})$, it suffices to prove that the following terms vanish for $k \to \infty$

$$0 \leq \left\{ \mathbb{E} \left[ (\hat{m}_k(X'(e_N)) - Y(e_N))^2 \mid X(I_{n,k}), Y(I_{n,k}) \right] \right\}^{1/2}$$

$$+ \left\{ \inf_{f \in \mathcal{F}} \mathbb{E} \left[ (f(X(e_N)) - Y(e_N))^2 \right] - \mathbb{E} \left[ (m(X(e_N)) - Y(e_N))^2 \right] \right\}^{1/2}$$

$$=: T_{1,k} + T_{2,k}. \quad (4.1)$$

The second term $T_{2,k}$ in (4.1) converges to zero in the mean (resp. a.s.): this follows immediately with the reverse triangle inequality and the denseness assumption on the function spaces from [2.1]; in the case of Haar wavelet spaces from [2.2] we need here that the sequence $(u_k : k \in \mathbb{N})$ converges to infinity if the distribution of the $X(s)$ is not bounded in order to guarantee the denseness.

The first term $T_{1,k}$ in (4.1) can be bounded in the following way (cf. again [Kohler 2003])

$$T_{1,k} \leq 2 \mathbb{E} \left[ (Y(e_N) - Y_L(e_N))^2 \right]^{1/2} + 2 \left\{ \frac{1}{|I_{n,k}|} \sum_{s \in I_{n,k}} (Y(s) - Y_L(s))^2 \right\}^{1/2} + \max_{\tilde{F} \in \tilde{J}} \text{pen}_d(\tilde{F})$$

$$+ 2 \sup_{f \in \mathcal{F}} \left\{ \frac{1}{|I_{n,k}|} \sum_{s \in I_{n,k}} |f(X(s)) - Y_L(s)|^2 \right\}^{1/2} - \left\{ \mathbb{E} \left[ (f(X(e_N)) - Y_L(e_N))^2 \right] \right\}^{1/2}. \quad (4.3)$$

Evidently, $\text{pen}_d(\tilde{F}) \in O(\lambda_k^2 \mathcal{K}^*) \to 0$ by assumption. For a.s. convergence of the entire term $T_{1,k}$, we need the ergodicity of the random field $\{Y(s) : s \in \mathcal{N}\}$. This is guaranteed if the random field $Y$ is strong mixing and stationary by Theorem B.4. Hence, we have a.s.

$$\left\{ \frac{1}{|I_{n,k}|} \sum_{s \in I_{n,k}} (Y(s) - Y_L(s))^2 \right\}^{1/2} \to \mathbb{E} \left[ (Y(e_N) - Y_L(e_N))^2 \right]^{1/2} \text{ as } k \to \infty$$

and $\mathbb{E} \left[ (Y(e_N) - Y_L(e_N))^2 \right]^{1/2} \to 0$ as $L \to \infty$ with dominated convergence. Consequently, it remains to show that

$$S_k := \sup_{f \in \mathcal{F}} \left\{ \frac{1}{|I_{n,k}|} \sum_{s \in I_{n,k}} |f(X(s)) - Y_L(s)|^2 \right\}^{1/2} - \left\{ \mathbb{E} \left[ |f(X(e_N)) - Y_L(e_N)|^2 \right] \right\}^{1/2} \to 0, \quad (4.4)$$
in the mean (resp. a.s.). For the convergence in the mean of equation (4.2), use the fact that \( (\sqrt{a} - \sqrt{b})^2 \leq |a-b| \), thus, together with Hölder’s inequality on probability spaces, the mean of \( S_k \) satisfies

\[
\mathbb{E} \left[ |S_k| \right] \leq \mathbb{E} \left[ \left( \frac{1}{|\mathcal{F}_k|} \sum_{x \in \mathcal{F}_k} |f(X(x)) - Y_L(x)|^2 \right) \left( \mathbb{E} \left[ |f(X(e_N)) - Y_L(e_N)|^2 \right] \right)^{1/2} \right],
\]

and apply Theorem A.6 to the RHS. In case of a.s.-convergence, use again the relation \( |\sqrt{a} - \sqrt{b}| \leq \sqrt{|a-b|} \) and the continuity of the square root function. Theorems A.6 and Theorem A.3 apply in this case, too: we have for the tail distribution for \( \varepsilon > 0 \) fix

\[
\Pr \left( \sum_{x \in \mathcal{F}_k} |f(X(x)) - Y_L(x)|^2 \left( \mathbb{E} \left[ |f(X(e_N)) - Y_L(e_N)|^2 \right] \right)^{1/2} > \varepsilon \right) \leq \tilde{A}_1 H_{\tilde{f}_k} \left( \frac{\varepsilon}{128\beta_k} \right) \exp \left\{ - \frac{\tilde{A}_2 \varepsilon \left( \prod_{i=1}^N n_i(k) \right)^{1/(N+1)}}{\beta_k^2 \prod_{i=1}^N \log n_i(k)} \right\}
\]

where in the last inequality we use that the vector space dimension of \( \mathcal{F}_k \) is at most \( K^* \). The constants \( \tilde{A}_1, A_i \) depend on the lattice dimension, the bound on the mixing coefficients and \( \varepsilon > 0 \). One finds that (4.2) convergences to zero in the mean if

\[
\beta_k^2 K^* \log \beta_k \prod_{i=1}^N \log n_i(k) \left( \prod_{i=1}^N n_i(k) \right)^{1/(N+1)} \to 0 \text{ as } k \to \infty.
\]

a.s.-convergence of the term in (4.2) follows with an application of the first Borel-Cantelli Lemma if additionally, for some positive \( \delta > 0 \)

\[
\beta_k^2 (\log k)^{1+\delta} \prod_{i=1}^N \log n_i(k) \left( \prod_{i=1}^N n_i(k) \right)^{1/(N+1)} \to 0 \text{ as } k \to \infty.
\]

\[\square\]

**Lemma 4.1** (Variant of Lemma 1 in Kohler [2008]). Let \( f \in \mathcal{F}_e \circ \pi \) for a partition \( \pi \in \prod_u \) for \( 1 \leq u \leq u_{\max} \). Then for fix \( \omega \in \Omega \) there are balanced wavelets \( f_j, \ldots, f_j \in \mathcal{F}_k \) which depend on this \( \omega \in \Omega \), such that \( \mathcal{F}_e \circ \pi = \{ f_j, \ldots, f_j \} \) in \( L^2(\mu_u) \) and \( v \leq |\pi| = (2^d - 1)(u - 1) + (2w_k)^d \).

**Proof.** The proof follows with induction on \( 1 \leq u \leq u_{\max} = 1 + (2w_k)^d (2 |\pi| + 1) / (2^d - 1) \) and the definition of the set systems \( \prod_u \). If \( u = 1 \), then \( \prod_1 \) only contains the partition

\[
\pi = [\sup \Phi_{j,y} : j \in A_{j,1} \} = \{ 2^{-j} y \} : y \in A_{j,1} \}.
\]

For the inductive step, \( u \to u + 1 \), let \( \pi \in \prod_{u+1} \) be a partition and \( \pi' \in \prod_u \) the corresponding predecessor partition which satisfies the relationship

\[
\pi = (\pi' \setminus \{ A \}) \cup \{ A_{u+1} \} : u_i \in [L, R], i = 1, \ldots, d \}.
\]

By construction, in \( L^2(\mu_u) \) the following equality is true

\[
\mathcal{F}_e \circ \pi = \{ A_{u+1} \} : u_i \in [L, R], i = 1, \ldots, d \}
\]

where the \( f_j' \) are the orthonormal balanced wavelets on \( A \) from our construction and \( \prod_u \) \( \mathcal{F}_e \circ \pi' \). By the inductive step, \( \mathcal{F}_e \circ \pi' = \{ f_j, \ldots, f_j \} \) for certain wavelets \( f_j \), from the constructed orthonormal system with \( v \leq |\pi| = (2^d - 1)(u - 1) + (2w_k)^d \). Hence, \( \mathcal{F}_e \circ \pi \) can be represented with \( (2^d - 1)u + (2w_k)^d \) elements as

\[
\mathcal{F}_e \circ \pi = \{ f_j, \ldots, f_j, f_j', \ldots, f_{2^{d-1}} \}.
\]

This finishes the proof. \(\square\)
We formulate the following extension of the Bernstein inequality of Theorem A.4, which we need in the proof of Theorem 2.2.

**Theorem 4.2** (Extension of Theorem A.4). Let \( \{Z(s) : s \in I\} \) be a strong mixing random field with \( \mathbb{E} [Z(s)] = 0 \) and \( \mathbb{E} [Z(s)^2] \leq \sigma^2 < \infty \). Furthermore, assume that the tail distribution is bounded by

\[
\mathbb{P}(|Z(s)| > z) \leq \kappa_0 \exp(-\kappa_1 z^\tau)
\]

for \( \kappa_0, \kappa_1, \tau > 0 \). Then, for any \( B > 0 \), we have with the notation from Theorem A.4

\[
\mathbb{P} \left( \sum_{s \in I_n} Z(s) > \epsilon \right) \leq \frac{12}{\epsilon^\tau \kappa_0 \kappa_1} \Gamma(1 - \tau, \kappa_1 B^\tau) |I_n| + 2 \exp \left( D_1 \sqrt{\mathbb{E} Z(s) \mathbb{E} Z(s)} |I_n| + \frac{1}{\mathbb{E} \sigma_0} \sqrt{|I_n|} \right)
\]

\[
\cdot \exp \left\{ -\frac{1}{\tau} \beta \epsilon \right\} \cdot \exp \left[ \frac{3}{2} \beta^2 \epsilon \left( \sigma^2 + 16D_2 \sigma^2 \alpha_p \right) |I_n| \right],
\]

where \( \Gamma \) denotes the upper incomplete gamma function.

**Proof.** We split each \( Z(s) \): choose an arbitrary bound \( B > 0 \) and define for \( s \in \mathbb{Z}^N \)

\[
Z(s)^0 = Z(s) - \min(Z(s), B) \geq 0, \quad Z(s)^* = Z(s) - \max(Z(s), -B) \leq 0
\]

and \( Z(s)^0 = \max(\min(Z(s), B), -B) \).

Then, \( Z(s) = Z(s)^0 + Z(s)^* + Z(s)^0 \) and \( 0 = \mathbb{E} [Z(s)] = \mathbb{E} [Z(s)^0] + \mathbb{E} [Z(s)^*] + \mathbb{E} [Z(s)^0] \). Thus,

\[
\mathbb{P} \left( \sum_{s \in I_n} Z(s) > \epsilon \right) = \mathbb{P} \left( \sum_{s \in I_n} Z(s) - \mathbb{E} [Z(s)] > \epsilon \right)
\]

\[
\leq \mathbb{P} \left( \sum_{s \in I_n} Z(s)^0 - \mathbb{E} [Z(s)^0] > \frac{\epsilon}{3} \right) + \mathbb{P} \left( \sum_{s \in I_n} Z(s)^* - \mathbb{E} [Z(s)^*] > \frac{\epsilon}{3} \right)
\]

\[
+ \mathbb{P} \left( \sum_{s \in I_n} Z(s)^0 - \mathbb{E} [Z(s)^0] > \frac{\epsilon}{3} \right)
\]

We treat each term of (4.4) separately. We consider first the first two terms. Using Chebyshev’s inequality we obtain

\[
\mathbb{P} \left( \sum_{s \in I_n} Z(s)^0 - \mathbb{E} [Z(s)^0] > \frac{\epsilon}{3} \right)
\]

\[
\leq \frac{3}{\epsilon} \mathbb{E} \left\| \sum_{s \in I_n} Z(s)^0 - \mathbb{E} [Z(s)^0] \right\| \leq \frac{6 |I_n|}{\epsilon} \mathbb{E} \left[ Z(s)^0 \right].
\]

Using the tail condition, we can estimate the expectation in (4.5) by

\[
\mathbb{E} [Z(s)^0] = \int_0^\infty \mathbb{P} \left( Z(s)^0 > z \right) dz
\]

\[
= \int_0^B \mathbb{P} \left( (Z(s) - B)1_{|Z(s)| \geq B} > z \right) dz = \int_B^\infty \mathbb{P} \left( Z(s) > z \right) dz
\]

\[
\leq \kappa_0 \int_B^\infty \exp(-\kappa_1 z^\tau) dz = \kappa_0 \int_1^{\frac{1}{\kappa_1}} \left( \frac{1}{\kappa_1} \right)^{1/\tau} y^{\frac{1}{\tau} - 1} e^{-y} dy
\]

\[
= \frac{\kappa_0}{\tau} \left( \frac{1}{\kappa_1} \right)^{1/\tau} \Gamma \left( 1, \kappa_1 B^\tau \right).
\]

Since \( \sigma(Z(s)^0 : s \in I) \leq \sigma(Z(s) : s \in I) \) for any \( I \subseteq \mathbb{Z}^N \), the mixing coefficient of the field \( \{Z(s)^0 : s \in \mathbb{Z}^N\} \) can be estimated by those of \( \{Z(s) : s \in \mathbb{Z}^N\} \). Furthermore, \( \text{Var}(Z(s)^0) \leq \sigma^2 \) and we can apply Theorem A.4 to the third term of (4.4), using that \( |Z(s)^0 - \mathbb{E} [Z(s)^0]| \leq 2B \). Hence,

\[
\mathbb{P} \left( \sum_{s \in I_n} Z(s)^0 - \mathbb{E} [Z(s)^0] > \frac{\epsilon}{3} \right)
\]
Lemma 4.5 \((\text{Large deviations for heteroscedastic noise})\)

independent sample for the regression problem from equations

\[ (\text{Large deviations of strong mixing samples from independent samples}) \]

such that

\[ \text{can bound} \]

random field fulfills Condition 1.4 \((a)\) and \((b)\), has zero means and satisfies the tail condition

not on

16 JOHANNES T. N. KREBS

Proof. Let \(g_1, \ldots, g_N\) be an \(e/4\)-covering of \(G\) with respect to the \(L^1\)-norm of the empirical measure induced by \((Z(I_n), Z(I_n)) \subseteq \mathbb{R}^d\).

Then

\[ \mathbb{P}\left( \sup_{g \in G} \frac{1}{|I_n|} \sum_{s \in I_n} g(Z(s)) - \frac{1}{|I_n|} \sum_{s \in I_n} g(Z'(s)) > \varepsilon \right) \leq A_1 H_{\mathcal{G}} \left( \frac{e}{4} \right) \exp \left( -A_2 \left( \prod_{i=1}^N n_i \right)^{1/(N+1)} \right) . \]

The proof follows now with an application of Proposition A.5. \(\square\)

**Lemma 4.4** \((\text{Modified version of Theorem 11.4 of Györfi et al. 2002})\).

Let \((X(i), Y(i)) : i = 1, \ldots, n\) be an independent sample from the regression problem from equations \((1.1)\). Assume that the regression function \(m\) is essentially bounded, \(\|m\|_{\infty} \leq B < \infty\), for \(B \geq 1\). Let \(f\) be a function class such that each element \(f: \mathbb{R}^d \to \mathbb{R}\) and \(\|f\|_{\infty} \leq B\). Then given that \(\alpha, \beta, \gamma > 0\) and \(0 < \delta \leq 1/2\)

\[ \mathbb{P}\left( \sup_{f \in F} \mathbb{E}\left[ |f(X(e_n)) - m(X(e_n))|^2 \right] - \frac{1}{|I_n|} \left| \mathbb{E}\left[ |f(X(s)) - m(X(s))|^2 \right] \right| > \delta \right) \]

\[ \geq \delta \left( \alpha + \beta + \mathbb{E}\left[ |f(X(e_n)) - m(X(e_n))|^2 \right] \right) \leq 14H_F \left( \frac{\beta \delta}{20B} \right) \exp \left( -\frac{\beta^2 (1-\delta) \alpha |I_n|}{214 (1 + \delta) B^4} \right) . \]

Proof. One can deduce the claim from the proof of Theorem 11.4 of Györfi et al. 2002. \(\square\)

**Lemma 4.5** \((\text{Large deviations for heteroscedastic noise})\).

Suppose that the random field \(\varepsilon = \{\varepsilon(s) : s \in \mathbb{Z}^d\}\) fulfills Condition 1.3 \((a)\) and \((b)\), has zero means and satisfies the tail condition

\[ \mathbb{P}(\varepsilon(s) > z) \leq k_0 \exp(-k_1 z^2) \text{ for constants } 0 < k_0, k_1, \tau < \infty. \]

Let the function class \(G\) fulfill Condition 1.2 \((a)\) for functions \(g : \mathbb{R}^d \to \mathbb{R}\) and \(\|g\|_{\infty} \leq B \geq 1\). Let \(\chi : \mathbb{R}^d \to \mathbb{R}_{+}\) be essentially bounded. Furthermore, let \(\{x_i : s \in I_n\}\) be points in \(\mathbb{R}^d\) where \(I_n = \{s : e_n \leq s \leq n\}\) and \(n \in \mathbb{N}\) such that \(\min_{1 \leq n} n_i \geq e^2\). Let \(K \subseteq \mathbb{R}_{+}\). Then for two constants \(A_1, A_2 \in \mathbb{R}_{+}\) which depend on \(N\), the bound on the mixing coefficients and the tail parameters \(k_0, k_1, \tau\), which are independent of \(B, K, \delta\) and \(n\)

\[ \mathbb{P}\left( \sup_{g \in G} \left| \frac{1}{|I_n|} \sum_{s \in I_n} \chi(x_s) \varepsilon(s) g(x_s) \right| > \varepsilon \right) \]

\[ \leq \inf_{D_{10}} \frac{\beta^2}{8K^2 B} \left( A_1 D_1^{-\tau} \delta^{-1} \exp\left( -A_2 D_1^\tau \right) + A_1 \exp\left( -A_2 \left( \prod_{i=1}^N n_i \right)^{1/(N+1)} \right) \right) \]

\[ + \inf_{D_{10}} \left( A_1 |\delta|_{\infty} K^2 D_2^{-\tau/2} \exp\left( -A_2 D_2^\tau \right) + A_1 \exp\left( -A_2 \left( \prod_{i=1}^N n_i \right)^{1/(N+1)} \right) \right) . \]

Proof. We use the extended Bernstein inequality for unbounded random variables from Theorem 4.2. Here we can bound \(\Gamma(1/\tau, c_1 B^\tau)\) by \(c_0(c_1 B^\tau)^{-1/\tau} \exp(-c_1 B^\tau)\) for a suitable constant \(c_0 \in \mathbb{R}_{+}\), which depends on \(\tau\) but not on \(B\) and on \(c_1\). Before we start with the main proof, we do some preparation: we apply Theorem 4.2 to a random field \(W\) which has zero means and fulfills the tail condition \(\mathbb{P}(\varepsilon(s) > z) \leq k_0 \exp(-k_1 z^2)\): there are
suitable constants $A_1, A_2 \in \mathbb{R}$, which only depend on $\kappa_0, \kappa_1, \tau$, the lattice dimension $N$ and the bound on the mixing coefficients but not on $n$ and $\delta$ such that

$$
\mathbb{P}\left( \sum_{s \in I_n} W(s) > \delta \mid I_n \right) \leq \inf_{D > 0} A_1 D^{1-\tau} \delta^{-1} \exp\left(-A_2 D^2\right) + A_1 \exp\left(-\frac{A_2 \delta \left(\prod_{i=1}^N n_i\right)^{1/(N+1)}}{D \prod_{i=1}^N n_i \log n_i}\right).
$$

Furthermore, let there be given an $\tilde{\delta}$-covering of $G$ w.r.t. the $L^1$-norm induced by the empirical measure $|I_n|^{-1} \sum_{s \in I_n} \delta_s$, which we denote by $(\mathbb{g}_1, \ldots, \mathbb{g}_{N^*})$, for some $N^* \in \mathbb{N}$. Then, any function $g$ in the $\tilde{\delta}$-neighborhood of a covering function $g_j$ satisfies

$$
\sqrt{|I_n|^{-1} \sum_{s \in I_n} |g(x_s) - g_j(x_s)|^2} \leq \sqrt{|I_n|^{-1} \sum_{s \in I_n} |g(x_s) - g_j(x_s)|^2} 2B \leq \sqrt{2B} \tilde{\delta}.
$$

I.e., $(\mathbb{g}_1, \ldots, \mathbb{g}_{N^*})$ is a $\sqrt{2B} \tilde{\delta}$-covering w.r.t. the 2-norm. This means the $\tilde{\delta}$-covering number w.r.t. the 2-norm is bounded by $H_\tilde{\delta}\left(\tilde{\delta}^2/2B\right)$. Let now $K \in \mathbb{R}$ be given, then the desired probability is bounded by:

$$
\mathbb{P}\left( \left|\sum_{s \in I_n} (\delta_s(x) \mathbb{g}(s) g(x_s)) \right| > \tilde{\delta}\right) \\
\leq \mathbb{P}\left( \sup_{g \in \mathcal{G}} \left| \frac{1}{|I_n|} \sum_{s \in I_n} \delta_s(x) \mathbb{g}(s) g(x_s) \right| > \tilde{\delta} \right) + \mathbb{P}\left( \left\|\mathbb{g}\right\|_\infty^2 \sum_{s \in I_n} \mathbb{g}(s)^2 \geq K^2 \right) + \mathbb{P}\left( |\mathbb{g}(s)|^2 |\sum_{s \in I_n} \mathbb{g}(s)^2 > K^2 | \right)
$$

(4.6)

Now, let there be given a $(\delta/(2K))^2/(2B)$-covering of $G$ with respect to the $L^1$-norm of the measure $|I_n|^{-1} \sum_{s \in I_n} \delta_s$, which is an $\delta/(2K)$-covering w.r.t. the corresponding 2-norm. Observe that the random field $\mathbb{g}^2 = (\mathbb{g}(s)^2 : s \in \mathbb{Z}^N)$ fulfills the tail condition with $\tau/2$. Furthermore,

$$
\mathbb{P}\left( |\mathbb{g}(s)\mathbb{g}(s) g(x_s)| > z \right) \leq \mathbb{P}\left( |\mathbb{g}(s)| > (|\mathbb{g}\|_\infty B)^{-1} z \right),
$$

so for these random variables the constants in tail condition changes somewhat. Altogether, we can bound (4.6) as follows: apply the $\delta/(2K)$-covering $(\mathbb{g}_1, \ldots, \mathbb{g}_{N^*})$ and the Cauchy-Schwarz inequality to the first term inside the first probability, then (4.6) is bounded by

$$
H_\mathbb{g} \left( \frac{\delta/(2K)^2}{2B} \right) \sup_{1 \leq j \leq N^*} \mathbb{P}\left( \left| \frac{1}{|I_n|} \sum_{s \in I_n} \delta_s(x) \mathbb{g}(s) g_j(x_s) \right| \geq \frac{\delta}{2} \right) + \mathbb{P}\left( \left| \sum_{s \in I_n} \mathbb{g}(s)^2 \right| \geq K^2 \left|\mathbb{g}\|_\infty^2 \right) \right)
$$

$$
\leq \inf_{D_1 > 0} H_\mathbb{g} \left( \frac{\delta^2}{8K^2 B} \right) \left\{ A_1 D_1^{1-\tau} \delta^{-1} \exp\left(-A_2 D_1^2\right) + A_1 \exp\left(-\frac{A_2 \delta \left(\prod_{i=1}^N n_i\right)^{1/(N+1)}}{D_1 \prod_{i=1}^N n_i \log n_i}\right) \right\}
$$

$$
+ \inf_{D_2 > 0} \left\{ A_1 \left|\mathbb{g}\|_\infty^2 K^2 D_2^{-\tau/2} \exp\left(-A_2 D_2^2/2\right) + A_1 \exp\left(-\frac{A_2 \left|\mathbb{g}\|_\infty^2 K^2 \left(\prod_{i=1}^N n_i\right)^{1/(N+1)}\right)}{D_2 \prod_{i=1}^N n_i \log n_i}\right) \right\}
$$

where the constants $A_1, A_2$ are independent of $B$, $K$, $n$, $\delta$ and the $D_1$ but depend on the lattice dimension, the bound on the mixing coefficients and the tail parameters $\kappa_0, \kappa_1$ and $\tau$. This finishes the proof. □

We can now prove Theorem 2.2

**Proof of Theorem 2.2** Let $(X(I_{nk}), Y(I_{nk}))$ be an i.i.d. ghost sample on an enlarged probability space with the same marginal distributions as the observations $(X(s), Y(s))$ for the given sequence of index sets $I_{nk}$ from Condition 1.4 (a) and (b) with $I = \mathbb{Z}^N$. Let the truncation sequence be given by $\beta_k \equiv B$. We define the empirical norms for a real valued function $f$ on $\mathbb{R}^d$

$$
\left\|f\right\|_k := \left\|I_{nk}\right|^{-1} \sum_{s \in I_{nk}} f(X(s))^2 \right\|^{1/2} \quad \text{and} \quad \left\|f\right\|_k^2 := \left\|I_{nk}\right|^{-1} \sum_{s \in I_{nk}} f(X(s))^2 \right\|^{1/2}.
$$
Additionally, write \( \| \cdot \| \) for the \( L^2(\mu_X) \)-norm: \( \| f \|^2 = \int_{\mathbb{R}^d} f^2 \, d\mu_X \). The \( L^2 \)-error can be decomposed in three terms with the help of the i.i.d. ghost sample
\[
\int_{\mathbb{R}^d} \| \hat{m}_k - m \|^2 \, d\mu_X = \left\{ \| \hat{m}_k - m \|^2 - 2 \left( \| \hat{m}_k - m \|_k^2 + \text{pen}_k(m_k) \right) \right\} + 2 \left\{ \| \hat{m}_k - m \|_k^2 + \text{pen}_k(m_k) \right\} + 2 \left\{ \| \hat{m}_k - m \|_k^2 \right\} =: T_{1,k} + T_{2,k} + T_{3,k}.
\]
We investigate the terms \( T_{1,k} \) separately. We start with \( T_{1,k} \); note that \( \hat{m}_k \in T_B F_k \), consequently,
\[
\mathbb{P} \left( T_{1,k} > t \right) = \mathbb{P} \left( \| \hat{m}_k - m \|^2 - 2 \left( \| \hat{m}_k - m \|_k^2 + \text{pen}_k(m_k) \right) > t \right) \leq \mathbb{P} \left\{ \sup_{j \in F_k} \| T_B f - m \|^2 - 2 \left( \| T_B f - m \|_k^2 + \text{pen}_k(f) \right) > t \right\} \leq \mathbb{P} \left\{ \exists f \in T_B F_k : \mathbb{E} \left[ (f(X(e_N)) - m(X(e_N)))^2 \right] \right\} \leq \frac{1}{2} \left( 1 + \mathbb{E} \left[ (f(X(e_N)) - m(X(e_N)))^2 \right] \right),
\]
(4.7)
here we can omit the penalizing term because \( \text{pen}_k(f) \geq 0 \). Apply Lemma 4.4 to equation (4.7) with the parameters \( \alpha = \beta = t/2 \) and \( \delta = 1/2 \):
\[
4.7 \leq 14 H_{T,B} F_k \left( \frac{t/2}{20B} \right) \exp \left( \frac{t/2 |I_{n(k)}|}{256B^4} \right) \leq C_1 \exp \left( C_2 K^* \log(B^2/t) - C_3 t |I_{n(k)}|/B^4 \right),
\]
where we use that both
\[
\log H_{T,B} F_k \left( \frac{t/2}{20B} \right) \leq C_1 \left( \text{V}(T_{B} F_k) \log \left( \frac{C_2 B^2}{t} \right) \right) \text{ and } \text{V}(T_{B} F_k) \leq K^* + 1.
\]
The constants \( C_1, C_2, C_3 \) do not depend on \( I_{n(k)}, K^*, t \) or \( B \). Hence, we choose \( v_k := K^* |I_{n(k)}|^{-1} \left( \prod_{i=1}^N \log n_i(k) \right)^2 \) and the expectation of the first term can be bounded by
\[
\mathbb{E} \left[ T_{1,k} \right] \leq v_k + \int_{v_k}^{\infty} \mathbb{P} \left( T_{1,k} > t \right) \, dt \in O \left( K^* \left( \prod_{i=1}^N \log n_i(k) \right)^2 |I_{n(k)}| \right).
\]
(4.8)
We study the second term \( T_{2,k} \); therefore define the function which minimizes the penalized sum of squares w.r.t. the true mean function \( m \)
\[
m_k^* := \arg \min_{m \in F_k} \left\{ \frac{1}{|I_{n(k)}|} \sum_{x \in I_{n(k)}} (f(X(s)) - m(X(s)))^2 + \text{pen}_k(f) \right\}.
\]
We compute the conditional expectation of \( T_{2,k} \) given the data \( X(I_{n(k)}) \) and use the pointwise inequality \( |\hat{m}_k - m| \leq |m_k - m| \) which is true because both \( \hat{m}_k \) and \( m \) are bounded by \( B \),
\[
\frac{1}{2} \mathbb{E} \left[ T_{2,k} | X(I_{n(k)}) \right] \leq \mathbb{E} \left[ |m_k - m|_k^2 + \text{pen}_k(m_k) | X(I_{n(k)}) \right] \leq \mathbb{E} \left[ |m_k - m|_k^2 + \text{pen}_k(m_k) - 2 \left( m_k^* - m \right)_k^2 + \text{pen}_k(m_k^*) \right] + 2 \left( |m_k^* - m|_k^2 + \text{pen}_k(m_k^*) \right) \leq v_k + \int_{v_k}^{\infty} \mathbb{P} \left( |m_k - m|_k^2 + \text{pen}_k(m_k) > 2 \left( m_k^* - m \right)_k^2 + \text{pen}_k(m_k^*) + 1 \right) \, dt \]
\[
+ 2 \inf_{f \in F_k} \left\{ \frac{1}{|I_{n(k)}|} \sum_{x \in I_{n(k)}} (f(X(s)) - m(X(s)))^2 + \text{pen}_k(f) \right\},
\]
(4.9)
for \( v_k > 0 \) and where we use the continuity properties of a conditional distribution function as well as the defining property of \( m_k^* \). Set

\[
V^2(m_k|m_k^*) := \left\| m_k - m_k^* \right\|^2 + \text{pen}_k(m_k)
\]

and consider the conditional distribution in equation (4.9): one can show with elementary calculations, cf. the proof of Theorem 2.1 in [van de Geer 2001] that by the definitions of \( m_k \) and \( m_k^* \) for given data \( X(I_{(k)}) = x(I_{(k)}) := \{ x_s : s \in I_{(k)} \} \subseteq \mathbb{R}^d \) the inclusion

\[
\left\{ \left\| m_k - m_k^* \right\|^2 + \text{pen}_k(m_k) > 2 \left( \left\| m_k^* - m_k \right\|^2 + \text{pen}_k(m_k^*) \right) + t \right\}
\]

\[
\subseteq \left\{ \frac{1}{|I_{(k)}|} \sum_{x \in I_{(k)}} \varsigma(x) \epsilon(x) \left( m_k(x) - m_k^*(x) \right) \geq V^2(m_k|m_k^*)/12 \text{ and } V^2(m_k|m_k^*) \geq 1 \right\}
\]

is true. Hence, the conditional distribution from equation (4.9) can be bounded as

\[
\mathbb{P}\left( \left\| m_k - m_k^* \right\|^2 + \text{pen}_k(m_k) > 2 \left( \left\| m_k^* - m_k \right\|^2 + \text{pen}_k(m_k^*) \right) + t \right)\bigg|_{X(I_{(k)}) = x(I_{(k)})} \leq \frac{1}{|I_{(k)}|} \sum_{x \in I_{(k)}} \varsigma(x) \epsilon(x) \left( m_k(x) - m_k^*(x) \right) \geq V^2(m_k|m_k^*)/12 \text{ and } V^2(m_k|m_k^*) \geq 1 \right\}
\]

\[
\leq \sum_{i=0}^\infty \mathbb{P}\left( \exists f \in \mathcal{F}_k : \frac{1}{|I_{(k)}|} \sum_{x \in I_{(k)}} \varsigma(x) \epsilon(x) \left( f(x) - m_k^*(x) \right) \geq \frac{2^{2i}}{12} \right)
\]

and

\[
V^2(m_k|m_k^*) \leq 2^{2(i+1)}t.
\]

Thus, it suffices to show that (4.10) can be bounded suitably. Define for \( \delta > 0 \) the functions classes

\[
\mathcal{G}_\delta(t) := T(\mathcal{I}(t)) \sqrt{\frac{2^i}{2^{i+1}}} \sqrt{t} \left\{ f - m_k^* : f \in \mathcal{F}_k, \ V^2(f|m_k^*) \leq 2^{2(i+1)}\delta \right\}.
\]

The function class \( \mathcal{G}_\delta(t) \) corresponds to the functions used in (4.10). Note that we can truncate the functions at \( \pm \sqrt{|I_{(k)}|} \sqrt{t} \) because the admissible \( f \in \mathcal{F}_k \) fulfill

\[
|f(X(x)) - m_k^*(X(x))| \leq \sqrt{|I_{(k)}|} \sqrt{t}.
\]

Next, set

\[
R(n) := |I_{(k)}|^{1/(N+1)} \prod_{i=1}^N \log n_i \text{ for } n \in \mathbb{N}^N.
\]

Now we are able to apply Lemma 4.3 to the probabilities in the sum in (4.10) (with \( K := 2\sqrt{t} \)). So, (4.10) equals

\[
\sum_{i=0}^\infty \mathbb{P}\left( \sup_{g \in \mathcal{G}_\delta(t)} \left\{ \frac{1}{|I_{(k)}|} \sum_{x \in I_{(k)}} \varsigma(x) \epsilon(x) g(x) \geq \frac{2^{2i}}{12} \right\} \right)\bigg|_{X(I_{(k)}) = x(I_{(k)})} \leq \sum_{i=0}^\infty \inf_D H_{\mathcal{G}_\delta(t)} \left( \frac{(2^{2i}/12)^2}{8 \cdot 2^{2i} |I_{(k)}|^{1/2} 2^{i+1} \sqrt{t}} \right)
\]

\[
\cdot \left\{ C_1 D_1^{1/2} \left( 2^{2i}/12 \right)^{-1} \exp \left( -C_2 D_1^{1/2} \right) + C_1 \exp \left( -C_2 D_1^{1/2} R(n) \right) \right\} + \sum_{i=0}^\infty \inf_D \left\{ C_1 D_2^{1/2} \left( 2^{2i}/12 \right)^{-1} \exp \left( -C_2 D_2^{1/2} \right) + C_1 \exp \left( -C_2 D_2^{1/2} R(n) \right) \right\},
\]
the constants \( C_1 \) and \( C_2 \) only depend on the lattice dimension \( N \), the bound on the mixing coefficients and the tail parameters \( \kappa_0, \kappa_1, \kappa_2 \). Note that this bound is deterministic. The covering number of this function classes \( \mathcal{G}_{k, \delta}(\delta) \) can be bounded with help of the Vapnik-Chervonenkis dimension of \( \mathcal{F}_k \)

\[
H_{G_{k, \delta}}\left( \frac{(2^2 \delta/12)^2}{8 \cdot 2^2 \delta |I_{\text{end}}|^{1/2} 2^{1+1} \sqrt{\delta}} \right) \leq H_{G_{k, \delta}}\left( \frac{2 \sqrt{5}}{2304 |I_{\text{end}}|^{1/2}} \right) \leq (C |I_{\text{end}}|)^{2V_{G_{k, \delta}}(\delta)} \leq (C |I_{\text{end}}|)^{2(K^*+2)}
\]

because \( \mathcal{G}_{k, \delta}(\delta) \subseteq T \sqrt{\ln(1/\delta)} \mathcal{F}_k, m_1^* \) and the linear space \( \mathcal{F}_k, m_1^* \) has a vector space dimension of at most \( K^* + 1 \); the bound can then be deduced from Proposition A.3.

Note that equation (4.12) is summable over the index \( l \) for all \( D_1, D_2 \in \mathbb{R}^+, \) which are independent of \( l \). We have again for suitable constants (which only depend on the lattice dimension \( N \), the bound on the mixing coefficients and the tail parameters)

\[
(4.12) \leq (C_1 |I_{\text{end}}|)^{2(K^*+2)} \cdot \inf_{D_1 \geq 0} \left\{ D_1^{-1} \exp\left( -C_2 D_1^2 \right) + \exp\left( -C_2 \frac{l R(n(k))}{D_1} \right) \right\} + C_1 \cdot \inf_{D_2 \geq 0} \left\{ D_2^{-1} \exp\left( -C_2 D_2^2 \right) + \exp\left( -C_2 \frac{l R(n(k))}{D_2} \right) \right\}. \tag{4.13}
\]

Set the parameter \( D_1 \) for each \( t \) such that the asymptotic growth rate of the two exponential terms are equal inside each factor of curvly brackets of (4.13), i.e.

\[
D_1 := t^{1/(1+r)} \left( |I_{\text{end}}|^{1/(N+1)} \prod_{i=1}^N \log n_i(k) \right)^{1/(r+1)} \quad \text{and} \quad D_2 := D_1^{1/(r)/(1+r/2)}.
\]

In particular, we find

\[
\int_{v_3}^{\infty} \exp\left( -C_2 (v_3/t)^{(1+r)} R(n(k))^{(1+r)} \right) \ dt \leq C \left( 1 + \frac{1}{\tau} \right)^{1/(1+r)} \tau \exp\left( -C_2 (v_3 R(n(k)))^{(1+r)} \right) R(n(k))^{(1+r)} \right) \tag{4.14}
\]

In addition, we have \( D_1^{-1} t^{-1} = t^{-2/(1+r)} R(n(k))^{(1+r)/(1+r)} \), hence, this factor is decreasing in \( t \). Define

\[
v_3 := K^* \left( \prod_{i=1}^N \log n_i(k) \right)^{2/(1+r)} R(n(k)).
\]

If we combine (4.13) with (4.14), we find that the integral from \( v_3 \) to \( \infty \) over the integrand in the first line of (4.13) decreases at a speed which is asymptotically in

\[
O \left( K^* \left( \prod_{i=1}^N \log n_i(k) \right)^{2/(1+r)} R(n(k)) \right).
\]

In the same way, by formally replacing \( \tau \) with \( \tau/2 \), one finds that the integral over the integrand in the second line in (4.13) is in \( O \left( K^* \left( \prod_{i=1}^N \log n_i(k) \right)^{2(r+\tau)/r} R(n(k)) \right) \). With this reduction, we can estimate the integral in equation (4.9) as

\[
\int_{v_2}^{\infty} \mathbb{P} \left[ \| m_k - m_0 \|_2^2 + pen(m_k) > 2 \left( \| m_k^* - m_0 \|_2^2 + pen(m_k^*) \right) + t \left| X(I_{\text{end}}) \right| \right] \ dt \leq C \left( K^* \left( \prod_{i=1}^N \log n_i(k) \right)^{2(r+\tau)/r} R(n(k)) \right).
\]

where the constant \( C \) only depends on the lattice dimension \( N \), the bound on the mixing coefficients and the tail parameters. Hence, the expectation of \( T_{2,k} \) is bounded by

\[
\mathbb{E} \left[ T_{2,k} \right] \leq 4 \mathbb{E} \left[ \inf_{f \in \mathcal{F}_k} \left\{ \frac{1}{|I_{\text{end}}|} \sum_{n \in I_{\text{end}}} \left| f(X(s)) - m(X(s)) \right|^2 + pen_k(f) \right\} \right] + C \left( K^* \left( \prod_{i=1}^N \log n_i(k) \right)^{2(r+\tau)/r} R(n(k)) \right)
\]

\[
\leq 4 \inf_{f \in \mathcal{F}_k} \left( \int_{\mathbb{R}^d} |f - m|^2 \, d\mu_X + \mathbb{E} \left[ pen_k(f) \right] \right) + C \left( K^* \left( \prod_{i=1}^N \log n_i(k) \right)^{2(r+\tau)/r} R(n(k)) \right). \tag{4.15}
\]
\[ 4 \min_{1 \leq i \leq K} \left\{ \inf_{f \in F_i} \int_{\mathbb{R}^d} |f - m|^2 \, d\mu_x + \alpha t_i^2 \right\} + C \left( \frac{(K^* \prod_{i=1}^N \log n_i(k))^2}{T(n(k))} \right). \]

Especially in the case of the wavelet system we can bound (4.15) slightly better, if we use Lemma 4.1

\[ 4.15 \quad 4 \min_{1 \leq i \leq n_{\text{max}}} \left\{ 2^{i/2} \left( \frac{2^d - 1}{2(2^d - 1)} \right) + \min_{m \in \mathcal{I}_x} \int_{\mathbb{R}^d} |f - m|^2 \, d\mu_x \right\} \]

where \( n_{\text{max}} = 1 + (2w_2)^d[(2^{d(j_i - j_0)} - 1)/(2^d - 1)] \) is the maximum index of the sets of partitions given in equation (2.5). We consider the third term. Define the function class \( G_\varepsilon := \{ g_f := (f - m)^2 : f \in \mathcal{T}_\varepsilon \} \). Let \( f_1, \ldots, f_s \) be an \( \varepsilon \)-cover of \( \mathcal{T}_\varepsilon \) w.r.t. the \( L^1 \)-norm of the empirical measure of the points \( (x_1, \ldots, x_s) \subseteq \mathbb{R}^d \). As both \( m \) and the functions \( f \) in \( \mathcal{T}_\varepsilon \) are bounded by \( B \), we have that the functions in \( G_\varepsilon \) are bounded by \( 4B^2 \). Furthermore, the functions \( g_f(x) := \|f(x) - m(x)\|^2 \) are a \( 4B\varepsilon \)-cover of \( G_\varepsilon \) w.r.t. the \( L^1 \)-norm of the empirical measure induced by \( x_1, \ldots, x_u \subseteq \mathbb{R}^d \). Indeed, let \( f \in \mathcal{T}_\varepsilon \) be in the neighborhood of \( f_j \) and denote by \( g_f \) resp. \( g_j \) the corresponding functions, then

\[ \frac{1}{u} \sum_{i=1}^u \left| \hat{g}_f(x_i) - \hat{g}_j(x_i) \right| = \frac{1}{u} \sum_{i=1}^u \left| \int (f(x_i) - m(x_i))^2 - (f_j(x_i) - m_j(x_i))^2 \right| \]

\[ \leq \frac{4B}{u} \sum_{i=1}^u \left| f(x_i) - f_j(x_i) \right| \leq 4B \varepsilon. \]

Consequently, \( H_{\mathcal{T}_\varepsilon}(t/4) \leq H_{\mathcal{T}_\varepsilon}(t/(16B)) \) and with Lemma 5.3, we obtain for the distribution of \( T_{3,k} \) the following inequalities

\[ \mathbb{P} \left( \|\hat{m}_k - m\|^2 - \|\hat{m}_k - m\|^2 > t \right) \]

\[ \leq \mathbb{P} \left( \sup_{f \in \mathcal{T}_{1/16B}} \left\{ \frac{1}{|U_{m,k}|} \sum_{x \in I_{m,k}} (f(X'(s)) - m(X'(s)))^2 \right\} > t \right) \]

\[ \leq \mathcal{H}_{U_{m,k}} \left( \frac{t}{16B} \right) \sup_j \mathbb{P} \left( \left\{ \frac{1}{|U_{m,k}|} \sum_{x \in I_{m,k}} (f_j(X'(s)) - m(X'(s)))^2 \right\} > \frac{t}{2} \right) \]

\[ \leq C_1 \exp \left( C_2 K^* \log(B^2/t) - C_3 R(n(k)) t/B^2 \right), \]

(4.16) for suitable constants \( C_1, C_2, C_3 \in \mathbb{R}_+ \) which only depend on the lattice dimension \( N \) and the bound on the mixing coefficients. Here, we use \( \mathcal{V}_{U_{m,k}} \leq K^* + 1 \). Hence, the expectation of the third term is bounded as

\[ \mathbb{E} \left[ T_{3,k} \right] \leq v_k + C_1 \int \exp \left( C_2 K^* \log(1/v_k) \right) \int_{v_k}^\infty \exp \left( -C_3 t R(n(k)) \right) \, dt \]

\[ \in O \left( K^* \left( \prod_{i=1}^N \log n_i(k) \right)^2 / R(n(k)) \right). \]

All in all, \( T_{1,k} \) and \( T_{3,k} \) are both negligible and the asymptotic properties are determined by the second term \( T_{2,k} \). \( \square \)

**Proof of Theorem 2.2** The proof can be carried out in the same way as the proof of Theorem 2.2. The bounds on the terms \( T_{1,k} \) and \( T_{3,k} \) do not change, both terms are in \( O \left( K^* \left( \prod_{i=1}^N \log n_i(k) \right)^2 / R(n(k)) \right) \). The second term can be treated in the same way until equation (4.11). Here use Theorem A.6 to obtain constants

\[ 4.6 \quad \sum_{i=0}^\infty C_i \left( \frac{C_2 (J_{m,k})^{2/(1+2)} \left( \frac{K^*}{2^d t} \right)^2}{2^d t} \right) \]

\[ \exp \left( -C_3 2^d t R(n(k)) \right) \]
\[ \leq C_1 \left( \frac{C_2 \sqrt{|I_{n(k)}|}}{\sqrt{N}} \right)^{2(K^*+2)} \exp \left( -C_3 tR(n(k)) \right). \]

With this bound it is straightforward to show
\[ \mathbb{E} \left[ T_{Z,1} \right] \leq 4 \inf_{f \in C} \left\{ \int_{\mathbb{R}^d} |f - m|^2 \, d\mu_x + \mathbb{E} \left[ \text{pen}_f(f) \right] \right\} + C \frac{K^* \left( \prod_{i=1}^{N} \log n_i(k) \right)^2}{R(n(k))} \]
and we are back in equation (4.15). In this case, the constant \( C \) only depends on the lattice dimension \( N \) and the bound on the mixing coefficients. This finishes the proof. \( \square \)

**Appendix A. Exponential inequalities for dependent sums**

We start with a definition of the covering number:

**Definition A.1 (\( \varepsilon \)-covering number).** Let \( (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d)) \) be endowed with a probability measure \( \nu \) and let \( \mathcal{G} \) be a set of real valued Borel functions on \( \mathbb{R}^d \) and let \( \varepsilon > 0 \). Every finite collection \( g_1, \ldots, g_N \) of Borel functions on \( \mathbb{R}^d \) is called an \( \varepsilon \)-cover of \( \mathcal{G} \) w.r.t. \( \| \cdot \|_{L^p(\nu)} \) of size \( N \) if for each \( g \in \mathcal{G} \) there is a \( j, 1 \leq j \leq N \), such that \( \| g - g_j \|_{L^p(\nu)} < \varepsilon \). The \( \varepsilon \)-covering number of \( \mathcal{G} \) w.r.t. \( \| \cdot \|_{L^p(\nu)} \) is defined as
\[ N(\varepsilon, \mathcal{G}, \| \cdot \|_{L^p(\nu)}) := \inf \{ N \in \mathbb{N} : \exists \varepsilon - \text{cover of } \mathcal{G} \text{ w.r.t. } \| \cdot \|_{L^p(\nu)} \text{ of size } N \}. \]

Evidently, the covering number is monotone:
\[ N(\varepsilon_2, \mathcal{G}, \| \cdot \|_{L^p(\nu)}) \leq N(\varepsilon_1, \mathcal{G}, \| \cdot \|_{L^p(\nu)}) \text{ if } \varepsilon_1 \leq \varepsilon_2. \]

The covering number can be bounded uniformly over all probability measures for a class of bounded functions under mild regularity conditions. Thus, the following covering condition is appropriate for many function classes \( \mathcal{G} \).

**Condition A.2 (Covering condition).** \( \mathcal{G} \) is a class of uniformly bounded, measurable functions \( f : \mathbb{R}^d \rightarrow \mathbb{R} \) such that \( \| f \|_{\infty} \leq B < \infty \) and for all \( \varepsilon > 0 \) and all \( N \geq 1 \) the following is true:

- For any choice \( z_1, \ldots, z_N \in \mathbb{R}^d \) the \( \varepsilon \)-covering number of \( \mathcal{G} \) w.r.t. the \( L^1 \)-norm of the discrete measure with point masses \( \frac{1}{N} \) in \( z_1, \ldots, z_N \) is bounded by a deterministic function depending only on \( \varepsilon \) and \( \mathcal{G} \), which we shall denote by \( H_{\mathcal{G}}(\varepsilon) \), i.e.,
\[ N(\varepsilon, \mathcal{G}, \frac{1}{N} \sum_{k=1}^{N} \delta_{z_k}) \leq H_{\mathcal{G}}(\varepsilon). \]

Denote by \( \mathcal{G}^* := \{(z,t) \in \mathbb{R}^d \times \mathbb{R} : t \leq g(z) : g \in \mathcal{G}\} \) the class of all subgraphs of the class \( \mathcal{G} \). Condition A.2 is satisfied if the Vapnik-Chervonenkis dimension of \( \mathcal{G}^* \) is at least two, i.e., \( \mathcal{V}_{\mathcal{G}^*} \geq 2 \) and if \( \varepsilon \) sufficiently small.

**Proposition A.3 (Bound on the covering number).** [Györfi et al. 2002] Theorem 9.4, [Haussler 1992]. Let \( [a, b] \subset \mathbb{R} \) be a finite interval. Let \( \mathcal{G} \) be a class of uniformly bounded real valued functions \( g : \mathbb{R}^d \mapsto [a, b] \) such that \( \mathcal{V}_{\mathcal{G}^*} \geq 2 \). Let \( 0 < \varepsilon < (b - a)/4 \). Then for any probability measure \( \nu \) on \( \mathcal{B}(\mathbb{R}^d) \)
\[ N(\varepsilon, \mathcal{G}, \| \cdot \|_{L^p(\nu)}) \leq 3 \left( \frac{2e(b-a)^p}{\varepsilon^p} \log \frac{3e(b-a)^p}{\varepsilon^p} \right)^{\mathcal{V}_{\mathcal{G}^*}}. \]

In particular, in the case that \( \mathcal{G} \) is an \( r \)-dimensional linear space, we have \( \mathcal{V}_{\mathcal{G}^*} \leq r + 1 \).

The Bernstein inequality from [Valenzuela-Domínguez and Franke 2005] from Theorem A.4 puts us in position to formulate the inequality which yields upper bounds on probability of the event of the type
\[ \left\{ \sup_{g \in \mathcal{G}} \left| \frac{1}{|I_s|} \sum_{s \in I_s} g(Z(s)) - \mathbb{E} \left[ g(Z(e_N)) \right] \right| > \varepsilon \right\}. \]

Of course, (A.1) is not an event for general function classes, however, we assume that the function classes in the present context are sufficiently regular such that (A.1) is \( \mathcal{A} \)-measurable.

**Theorem A.4 (Bernstein inequality for strong spatial mixing).** [Valenzuela-Domínguez and Franke 2005]. Let \( Z := \{ Z(s) : s \in I \} \) be a real-valued random field defined on a subset of the \( N \)-dimensional lattice \( \mathbb{Z}^N \). Let \( Z \) be strong mixing with mixing coefficients \( \{ \alpha_k : k \in \mathbb{N} \} \) such that each \( Z(s) \) is bounded by a uniform
constant $B$ and has expectation zero and the variance of $Z(s)$ is uniformly bounded by $\sigma^2$. Furthermore, put $\tilde{\alpha}_k := \sum_{n=1}^k d^n \alpha_n$. Then for all $\varepsilon > 0$ and $\beta > 0$ such that $0 < 2^{N+1} B \beta \varepsilon < 1$

$$\mathbb{P}\left( \left| \sum_{s \in D_k} Z(s) \right| > \varepsilon \right) \leq 2 \exp \left\{ D_1 \sqrt{\varepsilon} \frac{\tilde{n}}{P} \sqrt[4]{|\{i|2^n+1\}|} \right\} \cdot \exp \left\{ -\beta \varepsilon + 2^{3N} \beta^2 \varepsilon (\sigma^2 + 4D_2 B^2 \tilde{\alpha}_k) \tilde{n} \right\},$$

(A.2)

where $D_1, D_2 > 0$ are constants depending on the dimension $N$ and $P(n), Q(n)$ are arbitrary non-decreasing sequences in $\mathbb{N}_+$ satisfying for each $1 \leq i \leq N$

$$1 \leq Q(n_i) \leq P(n_i) < Q(n_i) + P(n_i) < n_i$$

and

$$\tilde{n} := n_1 \cdot \ldots \cdot n_N, \quad \tilde{P} := P_1(n_1) \cdot \ldots \cdot P_N(n_N)$$

$$q := \min \{ Q_1(n_1), \ldots, Q_N(n_N) \}, \quad P := \max \{ P_1(n_1), \ldots, P_N(n_N) \}.$$ 

We give two results which are immediate consequences of Theorems A.4 and 4.2.

**Proposition A.5.** Let the real valued random field $Z$ satisfy Condition [A.4]. The $Z(s)$ have expectation zero and are bounded by $B$. There are constants $A_1, A_2 \in \mathbb{R}_+$ which depend on the lattice dimension $N$ and the bound on the mixing coefficients but not on $n \in \mathbb{N}_+$ and not on $B$ such that for all $n \in \mathbb{N}_+$ with $\min_{1 \leq i \leq n} n_i \geq \varepsilon^2$ and $\varepsilon > 0$

$$\mathbb{P}\left( \left| \sum_{s \in D_k} Z(s) \right| > \varepsilon \right) \leq A_1 \exp \left\{ -A_2 \varepsilon B^{-1} \left( \prod_{i=1}^N n_i \right)^{-N/(N+1)} \left( \prod_{i=1}^N \log n_i \right)^{-1} \right\}.$$ 

**Proof of Proposition A.5.** We make the definitions: $P_i(n_i) := Q_i(n_i) := \left| n_i^{N/(N+1)} \log n_i \right|$ for $i = 1, \ldots, N$. Furthermore, we denote the smallest coordinate of $n \in \mathbb{N}_N$ by $n^* := \min_{1 \leq i \leq N} n_i$. We consider the first factor of the RHS of (A.2) and show that under the stated conditions we have

$$\sup \left\{ \exp \left( D_1 \sqrt{\varepsilon} \frac{\tilde{n}}{P} \sqrt[4]{|\{i|2^n+1\}|} \right) : n \in \mathbb{N}_N, n^* \geq \varepsilon^2 \right\} < \infty.$$ 

(A.3)

By assumption we have that $a(q) \leq c_1 \exp(-c_2 q)$, for two constants $c_1, c_2 \in \mathbb{R}_{\geq 0}$ and $q := \min_{1 \leq i \leq N} Q_i$. Therefore it suffices to show that

$$\log(\tilde{n}/\tilde{P}) - c_2/(2^N + 1) \tilde{P} \tilde{n} \rightarrow -\infty \text{ as } n^* \rightarrow \infty.$$ 

(A.4)

Note that for $a, b \geq 2$, we have $ab \geq a + b$. We make the definition $\eta := N/N + 1$. Let $n^* \geq \varepsilon^2$, then for any constant $c \in \mathbb{R}_+$

$$\log \left( \left( \prod_{i=1}^N n_i \right)^{-\eta} \left( \prod_{i=1}^N \log n_i \right)^{-1} \right) - c(n^*)^\eta \log n^* \left( \prod_{i=1}^N n_i \right)^{-\eta} \left( \prod_{i=1}^N \log n_i \right)^{-1}$$

$$\leq (N + 1)^{-1} \sum_{i=1}^N \log n_i - c(n^*)^{\eta + N(\eta - 1)} \left( \prod_{i=1}^N \log n_i \right)^{N(\eta - 1)}$$

$$\leq (N + 1)^{-1} \sum_{i=1}^N \log n_i - c \left( \prod_{i=1}^N \log n_i \right)^{N}$$

$$= (N + 1)^{-1} - c \log n^* \prod_{i=1}^N \log n_i \rightarrow -\infty \text{ as } n^* \rightarrow \infty.$$ 

This proves (A.4) and consequently, that (A.3) is finite. We come to the second term inside the second factor of (A.2). Define $\beta := (2^{N+1} \varepsilon B \tilde{P})^{-1}$ which fulfills the requirements of Theorem A.4. Then

$$\sup \left\{ 2^{3N} \beta^2 \varepsilon (\sigma^2 + 4D_2 B^2 \tilde{\alpha}_k) \tilde{n} : n \in \mathbb{N}_N, n^* \geq \varepsilon^2 \right\} < \infty.$$ 

(A.5)

This proves that $\mathbb{P}\left( \sum_{s \in D_k} Z(s) > \varepsilon \right) \leq A \exp \left( -\varepsilon/(2^{N+1} e B \tilde{P}) \right)$ for a constant $A \in \mathbb{R}_+$.

**Theorem A.6** (Large deviations for strong spatial mixing data). Let $Z$ be a random field on $(\Omega, \mathcal{A}, \mathbb{P})$ which satisfies Condition [4.2] and has equal marginal distributions. Let $\mathcal{G}$ be a set of measurable functions $g : \mathbb{R}^d \rightarrow \mathbb{R}$.
Thus, we can decompose
\[ \sup_{g \in \mathcal{G}} \left( \frac{1}{|I_n|} \sum_{x \in I_n} g(Z(s)) - \mathbb{E}[g(Z(e))] \right) \]
and on increasing function classes with some redundant constants where the constants \( A_1, A_2 \) and \( A_3 \) only depend on the lattice dimension \( N \) and on the bound on the mixing coefficients given by \( c_0, c_1 \in \mathbb{R} \) in Condition 1.2.

Since in practice, we shall use the bound given in Theorem A.6 on an increasing sequence \((n(k) : k \in \mathbb{N}) \subseteq \mathbb{N}^N\) and on increasing function classes \( \mathcal{G}_k \) whose essential bounds \( B_k \) increase with the size of the index sets \( I_{n(k)} \). It is possible to omit the first factor in the above theorem under certain conditions: let a sequence of function classes \( \mathcal{G}_k \) with bounds \( B_k \) and a sequence \( (\epsilon_k : k \in \mathbb{N}_1) \subseteq [0,1] \) be given such that

\[ \lim_{k \to \infty} \epsilon_k |I_{n(k)}| \left| \left| \frac{1}{|I_n|} \sum_{x \in I_n} g(Z(s)) - \mathbb{E}[g(Z(e))] \right| > \epsilon_k \right\} \leq A_1 H_{\mathcal{G}} \left( \frac{\epsilon_k}{32} \right) \exp \left( -\frac{A_2 \epsilon_k |I_{n(k)}|}{B_k \left( \prod_{i=1}^N n_i(k) \right)^{N/(N+1)} \prod_{i=1}^N \log n_i(k)} \right) \]

with new constants \( A_1, A_2 \in \mathbb{R}_+ \).

**Proof of Theorem A.6.** We assume the probability space to be endowed with the i.i.d. random variables \( Z'(s) \) for \( s \in I_n \) which have the same marginal laws as the \( Z(s) \). We put for shorthand

\[ S_n(g) := \frac{1}{|I_n|} \sum_{x \in I_n} g(Z(s)) \quad \text{and} \quad S'_n(g) := \frac{1}{|I_n|} \sum_{x \in I_n} g(Z'(s)). \]

Thus, we can decompose

\[ \sup_{g \in \mathcal{G}} \left[ S_n(g) - \mathbb{E}[g(Z(e))] \right] \]

\[ \leq \sup_{g \in \mathcal{G}} \left[ S_n(g) - S'_n(g) \right] + \sup_{g \in \mathcal{G}} \left[ S'_n(g) - \mathbb{E}[g(Z'(e))] \right] \]

Then apply Theorem 9.1 from Györfi et al. [2002] to second term on the right-hand side of (A.6), which is bounded by

\[ \mathbb{P} \left( \sup_{g \in \mathcal{G}} \left| S'_n(g) - \mathbb{E}[g(Z'(e))] \right| > \frac{\epsilon}{2} \right) \leq 8 H_{\mathcal{G}} \left( \frac{\epsilon}{16} \right) \exp \left( -\frac{|I_n| c_2}{512 B^2} \right). \]

To get a bound on the first term of the right-hand side of (A.6), we apply for fix \( \omega \in \Omega \) Condition A.2 to the set \( \{ Z(s, \omega), Z'(s, \omega) : s \in I_n \} \). Let \( g^*_n(\omega) \) for \( k = 1, \ldots, H^* := H_{\mathcal{G}} \left( \frac{\epsilon}{32} \right) \) be chosen as in Condition A.2 possibly with some redundant \( g^*_n(\omega) \) for \( \tilde{H}(\omega) < k \leq H^* \) where \( \tilde{H}(\omega) \) is the number of non-redundant functions. Note that \( H^* \) is deterministic. Define the random sets for \( k = 1, \ldots, H^* \) by

\[ U_k(\omega) := \left\{ g \in \mathcal{G} : \frac{1}{|I_n|} \sum_{x \in I_n} \left| g(Z(s, \omega)) - g^*_n(Z(s, \omega)) \right| + \left| g(Z'(s, \omega)) - g^*_n(Z'(s, \omega)) \right| < \frac{\epsilon}{32} \right\}, \]

note that some \( U_k(\omega) \) might be redundant for \( \tilde{H}(\omega) < k \leq H^* \). This implies that for each \( \omega \in \Omega \) we can write \( \mathcal{G} = U_1(\omega) \cup \ldots \cup U_{H^*}(\omega) \), consequently,

\[ \mathbb{P} \left( \sup_{g \in \mathcal{G}} \left| S_n(g) - S'_n(g) \right| > \frac{\epsilon}{2} \right) = \mathbb{P} \left( \max_{1 \leq k \leq H^*} \sup_{g \in U_k} \left| S_n(g) - S'_n(g) \right| > \frac{\epsilon}{2} \right). \]
In the following we suppress the $\omega$-wise notation; let now $g \in U_k$ be arbitrary but fixed, then

$$|S_n(g) - S'_n(g)| \leq \frac{\varepsilon}{32} + |S_n(g^*_k) - S'_n(g^*_k)|.$$  \hfill (A.9)

Thus, using equation (A.9), we get for each summand in (A.8)

$$\mathbb{P}\left(\sup_{g \in U_k} |S_n(g) - S'_n(g)| > \frac{\varepsilon}{2}\right) \leq \mathbb{P}\left(|S_n(g^*_k) - S'_n(g^*_k)| > \frac{7\varepsilon}{16}\right)$$

$$\leq \mathbb{P}\left(|S_n(g^*_k) - \mathbb{E}\left[g^*_k(Z(e_N))\right]| > \frac{7\varepsilon}{32}\right) + \mathbb{P}\left(|S'_n(g^*_k) - \mathbb{E}\left[g^*_k(Z'(e_N))\right]| > \frac{7\varepsilon}{32}\right).$$  \hfill (A.10)

The second term on the right-hand side of (A.10) can be estimated using Hoeffding’s inequality, we have

$$\mathbb{P}\left(|S'_n(g^*_k) - \mathbb{E}\left[g^*_k(Z'(e_N))\right]| > \frac{7\varepsilon}{32}\right) \leq 2\exp\left\{-\frac{322}{98}B^2\right\}.\hfill (A.11)$$

We apply the Bernstein inequality for strong spatial mixing data from Theorem A.4 to the first term of equation (A.10). We obtain for the first term on the right-hand side of (A.10) with Proposition A.5

$$\mathbb{P}\left(|S_n(g^*_k) - \mathbb{E}\left[g^*_k(Z(e_N))\right]| > \frac{7\varepsilon}{32}\right) \leq 2\lambda 1 \exp\left(\frac{A2\varepsilon |I_n|}{B \prod_{i=1}^n n_i^{N/(N+1)} \prod_{i=1}^n \log n_i}\right).$$  \hfill (A.12)

And all in all, using that $H_B\left(\frac{\varepsilon}{16}\right) \leq H_B\left(\frac{\varepsilon}{32}\right)$ and with the help of equation (A.7), and equations (A.11) and (A.12) plugged in (A.10) and that again in (A.8) we get the result - using the notation $n = \prod_{i=1}^n n_i$.

$$\mathbb{P}\left(1 \int_{x \in G} g(Z(x)) - \mathbb{E}\left[g(Z(e_N))\right] > \varepsilon\right)$$

$$\leq 8H_B\left(\frac{\varepsilon}{16}\right) \exp\left(-\frac{\varepsilon^2 |I_n|}{512B^2}\right) + 2H_B\left(\frac{\varepsilon}{32}\right) \exp\left(-\frac{98\varepsilon^2 |I_n|}{322B^2}\right) + A1 \exp\left(-\frac{A2\varepsilon |I_n|}{B \prod_{i=1}^n n_i^{N/(N+1)} \prod_{i=1}^n \log n_i}\right).$$

This finishes the proof. \hfill \Box

**Appendix B. Ergodic theory for spatial processes**

In the next lines, we give a review on important concepts of ergodicity when dealing with random fields on subgroups of the discrete group $\mathbb{Z}^N$. For further reading consult Tempelman (2010).

**Definition B.1** (Dynamical systems and ergodicity). Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and $(G, +)$ a locally compact, abelian Hausdorff group which fulfills the second axiom of countability. We write for $x, y \in G$ arbitrary $x - y$ for $x + (-y)$ and $-y$ is the $+$-inverse of $y$. Furthermore, let $\nu$ be a Haar measure on $B(G)$, i.e. for all $x \in G$ and for all Borel sets $B \in \mathcal{B}(G)$ we have $\nu(B) = \nu(x + B)$.

A family of bijective mappings $\{T_x : \Omega \to \Omega, x \in G\}$ is called a flow if it fulfills the following three conditions

1. $T_x$ is measure-preserving, i.e. $\mathbb{P}(A) = \mathbb{P}(T_x A)$ for all $A \in \mathcal{A}$ and for all $x \in G$.
2. $T_{x + y} = T_x \circ T_y$ and $T_x \circ T_{-x} = I_G$ for all $x, y \in G$.
3. the map $G \times \Omega \ni (x, \omega) \mapsto T_x \omega$ is measurable $[\mathcal{B}(G) \otimes \mathcal{A} | \mathcal{A}]$.

Let $T = \{T_x : x \in G\}$ be a flow in $(\Omega, \mathcal{A}, \mathbb{P})$, then the quadruple $(\Omega, \mathcal{A}, \mathbb{P}, T)$ is called a dynamical system. The dynamical system is called ergodic if the invariant $\sigma$-field $I := \{A \in \mathcal{A} : A = T_x A \forall x \in G\}$ is trivial $[\mathbb{P}]$, i.e. if for all $A \in I$ we have $\mathbb{P}(A) \in \{0, 1\}$.

Let now $\Gamma \leq \mathbb{Z}^N$ be a subgroup and $Z = \{Z(s) : s \in \Gamma\}$ be a stationary random field on $(\Omega, \mathcal{A}, \mathbb{P})$ where each $Z(s)$ takes values in the measure space $(S, \mathcal{S})$. Let $\nu$ be the counting measure on $\mathcal{B}(\Gamma)$. Put $\mathbb{P}_z := \mathbb{P}_{\{Z(s) : s \in \Gamma\}}$ for the probability measure on $\otimes_{\text{ref}} \mathfrak{S}$ induced by the finite dimensional distributions of $Z$ and define on the path space $(\otimes_{\text{ref}} \mathfrak{S}, \otimes_{\text{ref}} \mathfrak{S}, \mathbb{P}_z) \otimes_{\text{ref}} \mathfrak{E}$ the family of translations

$$T_t : \otimes_{\text{ref}} \mathfrak{S} \to \otimes_{\text{ref}} \mathfrak{S}, (z(s) : s \in \Gamma) \mapsto (z(s + t) : s \in \Gamma) \quad \text{for } t \in \Gamma,$$
which is a flow because \( Z \) is stationary. Then \( Z \) is called ergodic if and only if the quadruple \((X,\mathcal{F},\Theta,\mathbb{P})\) is ergodic.

The next result is an extension of Birkhoff's celebrated ergodic theorem it can be found in Tempelman [2010] Chapter 6, in particular Proposition 1.3 and Corrolary 3.2.

**Proposition B.3 (Stationarity and mixing imply ergodicity).** Let \( 0 \neq \Gamma \leq \mathbb{Z}^N \) be a subgroup and let the probability space \((\Omega,\mathcal{A},\mathbb{P})\) be endowed with the stationary process \( Z = \{Z(s) : s \in \Gamma \} \) for which each \( Z(s) \) takes values in \((S,\pi)\) and which fulfills the strong mixing condition from Definition [1.3]. Then \( Z \) is ergodic.

**Proof.** Let \( A \in \mathcal{I} \) be an \( T \)-invariant set of paths of \( Z \), it suffices to show that \( \mathbb{P}(A) \in [0,1], \) i.e.

\[
\mathbb{P}_Z(A) = \mathbb{P}(z \in A \cap T_s A) \to \mathbb{P}(z \in A)\mathbb{P}_Z(T_s A) = \mathbb{P}_Z(A)^2 \text{ as } x \to \infty.
\]

Let \( \varepsilon > 0 \) be given and let \( A, B \in \Theta \) be two sets of paths of \( Z \). Then by Carathéodory's extension theorem there are \( m,n \in \mathbb{Z} \) such that there are \( A^m \in \Theta, B^m \in \Theta \) with the property that both

\[
\mathbb{P}_Z(A \Delta A^m) < \frac{\varepsilon}{5} \text{ and } \mathbb{P}_Z(B \Delta B^m) < \frac{\varepsilon}{5}.
\]

Furthermore, by the strong mixing property from Definition [1.3] there is an \( x^* = r \cdot \varepsilon N \in \mathbb{Z}^N \) such that for \( x \geq x^*, x \in \Gamma \) we have

\[
|\mathbb{P}_Z(A^m \cap T_s B^m) - \mathbb{P}_Z(A^m)\mathbb{P}_Z(T_s B^m)| < \frac{\varepsilon}{5}.
\]

Consequently, we have for all \( x \geq x^* \)

\[
\left| \mathbb{P}(Z \in A, Z \in T_s B) - \mathbb{P}(Z \in A) \mathbb{P}(Z \in T_s B) \right| \\
\leq \left| \mathbb{P}(Z \in A \setminus A^m, Z \in T_s B) + \mathbb{P}(Z \in A^m, Z \in T_s B \setminus B^m) \right| \\
+ \left| \mathbb{P}(Z \in A^m, Z \in T_s B^m) - \mathbb{P}(Z \in A^m) \mathbb{P}(Z \in T_s B^m) \right| \\
+ \mathbb{P}(Z \in A^m) \mathbb{P}(Z \in T_s B \setminus B^m) + \mathbb{P}(Z \in A \setminus A^m) \mathbb{P}(Z \in T_s B) < \varepsilon.
\]

The main result in this section is the following one which generalizes Birkhoff's one dimensional ergodic theorem

**Theorem B.4.** Let \( 0 \neq \Gamma \leq \mathbb{Z}^N \) be a nontrivial subgroup and \( \{Z(s) : s \in \Gamma \} \) be a homogeneous strong mixing random field on \((\Omega,\mathcal{A},\mathbb{P})\) for some dimension \( N \in \mathbb{N} \). Let \( (n(k) : k \in \mathbb{N}) \subseteq \mathbb{N}^N \) be an increasing sequence such that \( \varepsilon N \leq n(k) \leq n(k+1) \) for which at least one coordinate converges to infinity. Then the sequence of index sets \( I_{n(k)} := \{ z \in \Gamma : \varepsilon N \leq z \leq n(k) \} \) is admissible in the sense of Theorem [1.3]. In particular, we have

\[
\frac{1}{|I_{n(k)}|} \sum_{z \in I_{n(k)}} Z(s) \to \mathbb{E}[Z(\varepsilon N)] \text{ a.s. as } k \to \infty.
\]

**Proof.** Since any subgroup of \( \mathbb{Z}^N \) is isomorphic to \( \mathbb{Z}^u \) for \( 0 \leq u \leq N, u \in \mathbb{N} \), it suffices to consider the case \( \Gamma = \mathbb{Z}^N, N \in \mathbb{N} \). In this case one computes easily that the regularity conditions of Theorem [B.3] are satisfied. The conclusion follows then from this theorem in combination with Proposition [B.3].
APPENDIX C. SIMULATION CONCEPTS

This section introduces an algorithm to simulate Markov random fields that are defined on arbitrary graphs \( G = (V, E) \) with a finite set of nodes \( V \). The main idea dates back at least to \cite{Kaiser2012} and is based on the concept of concliques which has the advantage that simulations can be performed faster when compared to the Gibbs sampler; an introduction to Gibbs sampling offers \cite{Bremaud1999}. We start with a definition

**Definition C.1** (Concliques, cf. \cite{Kaiser2012}). Let \( G = (V, E) \) be an undirected graph with a countable set of nodes \( V \) and let \( C \subseteq V \). If for all pairs of nodes \( (v, w) \in C \times C \) satisfying \( (v, w) \notin E \), the set \( C \) is called a conclique. A collection \( C_1, \ldots, C_n \) of concliques that partition \( V \) is called a conclique cover; the collection is a minimal conclique cover if it contains the smallest number of concliques needed to partition \( V \).

**Definition C.2** (Full conditional distribution). Let \( (\Omega, \mathcal{A}, \mathbb{P}) \) be a probability space and let \( (S, \mathcal{S}) \) be a state space. Let \( Y = (Y(s) : s \in I) \) be a collection of \( S \)-valued random variables. Then we call the family \( \{ \mathbb{P}(Y(s) \mid Y(t), t \in I \setminus \{s\}) \} \) a full conditional distribution of \( Y \).

Let now \( G \) be a graph whose nodes are partitioned into a conclique cover \( C_1, \ldots, C_n \). Let \( Y = (Y(v) : v \in V) \) be a Markov random field on \( G \) which takes values in \( (S, \mathcal{S}) \) with a full conditional distribution \( \{ F_v(Y(v) \mid Y(w), w \in Ne(v)) : v \in V \} \) and an initial distribution \( \mu_0 \). Note that the joint conditional distribution of a conclique \( Y(C_i) \) given its neighbors which are contained in \( Y(C_1), \ldots, Y(C_{i-1}), Y(C_{i+1}), \ldots, Y(C_n) \) factorizes as the product of the single conditional distributions due to the Markov property. This entails that we can — under mild regularity conditions — simulate the stationary distribution of the MRF with a Markov chain using the following algorithm:

**Algorithm C.3** (Simulation of random fields with concliques, \cite{Kaiser2012}). Simulate the starting values according to an initial distribution \( \mu_0 \) and obtain the vector of \( Y^{(0)} = (Y^{(0)}(C_1), \ldots, Y^{(0)}(C_n)) \). In the next step, given a vector \( Y^{(k)} = (Y^{(k)}(C_1), \ldots, Y^{(k)}(C_n)) \), simulate for \( i = 1, \ldots, n \) the concliques \( Y^{(k+1)}(C_i) \) given the \((k + 1)\)-st simulation of the neighbors in \( Y^{(k+1)}(C_1), \ldots, Y^{(k+1)}(C_{i-1}), Y^{(k+1)}(C_{i+1}), \ldots, Y^{(k+1)}(C_n) \). Then the transition kernel which captures the evolution of \( Y(C_i) \) given \( Y(C_{-i}) \) is given by

\[
\mathbb{M}_i : \ S^{|C_i| \times \mathcal{S}^{|C_i|}} \times [0, 1] \rightarrow [0, 1],
\]

\[
(y(C_{-i}), B) \mapsto \int \int_{\mathcal{S}^{C_i}} \int_{\mathcal{S}^{C_i}} M_i(y(C_{-i}), B) \nu^{C_i}(y(C_{-i}, y(\text{Ne}(i, s)))) \nu^{C_i}(y(C_{-i})) dy(C_i).
\]

With the help of \( \mathbb{M} \) the Markov kernel for the entire chain \( Y^{(k)} : k \in \mathbb{N} \) can be written as

\[
\mathbb{M} : \ S^{|V|} \times \mathcal{S}^{|V|} \rightarrow [0, 1],
\]

\[
(y, B) \mapsto \int \int_{\mathcal{S}^{C_1}} \int \int_{\mathcal{S}^{C_1}} M_1(y(C_{-1}), B) \mu_0 \left( (x(C_1), y(C_{-1}, 2)) \right) dx(C_1) \ldots
\]

\[
\ldots \int \int_{\mathcal{S}^{C_n}} \int \int_{\mathcal{S}^{C_n}} M_n(x(C_{-n}), B) \mu_0 \left( (x(C_{-1}), \ldots, x(C_{-n})) \right) dx(C_{-n}) \ldots
\]

\[
\ldots \int \int_{\mathcal{S}^{C_1}} \int \int_{\mathcal{S}^{C_1}} M_1(x(C_1), y(C_{-1}, 2)) \mu_0 \left( (x(C_1), y(C_{-1})) \right) dx(C_1) \ldots
\]

\[
\ldots \int \int_{\mathcal{S}^{C_n}} \int \int_{\mathcal{S}^{C_n}} M_n(x(C_{-n}), y(C_{-n})) \mu_0 \left( (x(C_{-1}), \ldots, x(C_{-n})) \right) dx(C_{-n}) \ldots.
\]

We are able to prove with these definitions

**Theorem C.4.** Let the density \( f \) be strictly positive on \( S^{|V|} \) such that the conditional densities \( f_{(C_i), y(\text{Ne}(i, s))} \) furnish a full conditional distribution, then the distribution of \( Y \), \( \mathbb{P}_Y \), is an invariant probability measure of the Markov chain given by equations \( \text{(C.1)} \) and \( \text{(C.2)} \) in the sense that \( \mathbb{P}_Y \mathbb{M} = \mathbb{P}_Y \). That is \( \mathbb{M} \) is positive.

It remains to prove the accuracy of the simulation approach of the homogeneous Markov chain simulated from a Markov random field as proposed in Algorithm \( \text{C.3} \) and equations \( \text{(C.1)} \) and \( \text{(C.2)} \) in the case that \( (S, \mathcal{S}) \subseteq \)
This means, we ask whether the chain is ergodic in the sense that \( \lim_{m \to \infty} \| \nu_m \|_{\mathfrak{M}} - \mathbb{P} \|_{\mathfrak{M}} = 0 \) in the total variation norm for the positive Markov kernel \( \mathfrak{M} \) with invariant probability measure \( \mathbb{P} \) and for all distributions \( \nu_0 \) on \( \mathfrak{M} \).

**Theorem C.5.** Let the Markov kernel \( \mathfrak{M} \) be given by equations (C.1) and (C.2) for the case that \( (S, \mathcal{E}) \subseteq (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d)) \). Assume that \( \mathfrak{M} \) arises from a full conditional distribution that is derived from a strictly positive joint density \( f \) w.r.t. the Lebesgue measure \( \lambda^d \). Then the Markov kernel is ergodic.

**Proof.** It suffices to verify that the requirements of the Aperiodic-Ergodic-Theorem are fulfilled, cf. Meyn and Tweedie [2009] Theorem 13.0.1. Plainly, the Markov kernel is \( \lambda^d \)-irreducible and \( \lambda^d \)-equivalent to any maximal irreducibility measure. Furthermore, since \( f \) is strictly positive, for any \( B \in \mathfrak{M} \) with positive Lebesgue measure, \( \mathfrak{M}(x, B) > 0 \) for all \( x \in S \). Hence, \( \mathfrak{M} \) is aperiodic. By Theorem C.4 the existence an invariant probability measure is fulfilled. By Theorem 10.1.1 and 10.0.1 in Meyn and Tweedie [2009] this invariant probability measure is unique. Furthermore, for each \( x \in S \) the probability measure \( \mathfrak{M}(x, \cdot) \) is absolutely continuous with respect to the Lebesgue measure \( \lambda^d \) which again is equivalent to the stationary measure \( \mathbb{P} = \int f \, d\lambda^d \) on \( S \). Thus, the requirements of Theorem 1.3 from Hernández-Lerma and Lasserre [2001] are met and the Markov chain in positive Harris recurrent and we can conclude from the Aperiodic-Ergodic-Theorem that \( \mathfrak{M} \) is ergodic. \( \square \)

We give an example

**Example C.6** (Concliques and the normal distribution). Let \( G = (V, E) \) be a finite graph and \( \{ Y(v) : v \in V \} \) be multivariate normal with expectation \( \alpha \in \mathbb{R}^{|V|} \) and covariance \( \Sigma \in \mathbb{R}^{|V| \times |V|} \) in that \( Y \) has the density

\[
f_Y(y) = (2\pi)^{-\frac{d}{2}} \det(\Sigma)^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} (y - \alpha)^T \Sigma^{-1} (y - \alpha) \right\}.
\]

Then for a node \( v \in V \) we use the notation \( P \) for the precision matrix \( \Sigma^{-1} \)

\[
Y(v) | Y(\cdot \backslash v) \sim \mathcal{N} \left( \alpha(v) - (P(v, v))^{-1} \sum_{w \neq v} P(w, v) (Y(w) - \alpha(w)), (P(v, v))^{-1} \right).
\]

Since \( P = \Sigma^{-1} \) is symmetric and since we can assume that \( (P(v, v))^{-1} > 0 \), \( Y \) is a Markov random field if and only if for all nodes \( v \in V \)

\[
P(v, w) \neq 0 \text{ for all } w \in Ne(v) \text{ and } P(v, v) = 0 \text{ for all } w \in V \setminus Ne(v).
\]

Cressie [1993] investigates the conditional specification

\[
Y(v) | Y(\cdot \backslash v) \sim \mathcal{N} \left( \alpha(v) + \sum_{w \in Ne(v)} c(v, w) (Y(w) - \alpha(w)), \tau^2(v) \right)
\]

where \( C = (c(v, w))_{v,w} \) is a \( |V| \times |V| \) matrix and \( T = \text{diag}(\tau^2(v) : v \in V) \) is a diagonal matrix such that the coefficients satisfy the necessary condition \( \tau^2(v)c(w, v) = \tau^2(w)c(v, w) \) for \( v \neq w \) and \( c(v, v) = 0 \) as well as \( c(v, w) = 0 \) if \( v, w \) are no neighbors. This means \( P(v, w) = -c(v, w)P(v, v) \), i.e. \( \Sigma^{-1} = P = T^{-1}(I - C) \).

If \( I - C \) is invertible and \( (I - C)^{-1}T \) is symmetric and positive definite, then the entire random field is multivariate normal with \( Y \sim \mathcal{N} \left( \alpha, (I - C)^{-1}T \right) \).

With this insight it is possible to simulate a Gaussian Markov random field using concliques with a consistent full conditional distribution. In particular, it is plausible in many applications to use equal weights \( c(v, w) \) (cf. Cressie [1993]): we can write the matrix \( C \) as \( C = \eta H \) where \( H \) is the adjacency matrix of \( G \), i.e. \( H(v, w) \) is 1 if \( v, w \) are neighbors, otherwise it is 0. We know from the properties of the Neumann series that \( I - C \) is invertible if \( (\eta h_m)^{-1} < \eta < (\eta h_0)^{-1} \) where \( h_0 \) is the minimal and \( h_m \) the maximal eigenvalue of \( H \).

**References**

J.J. Benedetto. *Wavelets: Mathematics and Applications*. Studies in Advanced Mathematics. Taylor & Francis, 1993.

Pierre Brémaud. *Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues*. Texts in Applied Mathematics. Springer, 1999.

N.A.C. Cressie. *Statistics for spatial data*. Wiley series in probability and mathematical statistics: Applied probability and statistics. J. Wiley, 1993.
V. Delouille, J. Franke, and R. von Sachs. Nonparametric stochastic regression with design-adapted wavelets. *Sankhya: The Indian Journal of Statistics. Special issue on Wavelets, Series A*, 63(3):328–366, 2001.

Maria Girardi and Wim Sweldens. A new class of unbalanced haar wavelets that form an unconditional basis for $l_p$ on general measure spaces. *Journal of Fourier Analysis and Applications*, 3(4):457–474, 1997. doi: 10.1007/BF02649107.

László Györfi, Michael Kohler, Adam Krzyżak, and Harro Walk. *A distribution-free theory of nonparametric regression*. Springer Berlin, New York, Heidelberg, 2002.

E.J. Hannan. *Multiple Time Series*. Wiley Series in Probability and Statistics. Wiley, 2009.

David Haussler. Decision theoretic generalizations of the pac model for neural net and other learning applications. *Information and computation*, 100(1):78–150, 1992.

Onésimo Hernández-Lerma and Jean B. Lasserre. Further criteria for positive Harris recurrence of Markov chains. *Proceedings of the American Mathematical Society*, 129(5):pp. 1521–1524, 2001.

Mark S. Kaiser, Soumendra N. Lahiri, and Daniel J. Nordman. Goodness of fit tests for a class of Markov random field models. *Ann. Statist.*, 40(1):104–130, 02 2012. doi: 10.1214/11-AOS948.

Michael Kohler. Nonlinear orthogonal series estimates for random design regression. *Journal of statistical planning and inference*, 115(2):491–520, 2003.

Michael Kohler. Multivariate orthogonal series estimates for random design regression. *Journal of Statistical Planning and Inference*, 138(10):3217 – 3237, 2008. doi: http://dx.doi.org/10.1016/j.jspi.2008.01.011.

Sean P. Meyn and Richard L. Tweedie. *Markov chains and stochastic stability*. Cambridge university press, 2009.

A.A. Tempelman. *Ergodic Theorems for Group Actions: Informational and Thermodynamical Aspects*. Mathematics and Its Applications. Springer Netherlands, 2010.

Eduardo Valenzuela-Domínguez and Jürgen Franke. A Bernstein inequality for strongly mixing spatial random processes. Technical report, Preprint series of the DFG priority program 1114 “Mathematical methods for time series analysis and digital image processing”, January 2005.

Sara van de Geer. Least squares estimation with complexity penalties. *Mathematical Methods of statistics*, 10 (3):355–374, 2001.

E-mail address: krebs@mathematik.uni-kl.de

University of Kaiserslautern, Erwin-Schrödinger-Strasse, 67663 Kaiserslautern