

On Integrated L^1 Convergence Rate of an Isotonic Regression Estimator for Multivariate Observations

Konstantinos Fokianos
University of Cyprus
Department of Mathematics and Statistics
P.O. Box 20537
CY – 1678 Nicosia
Cyprus
E-mail: fokianos@ucy.ac.cy

Anne Leucht
University of Bamberg
Department of Statistics and Mathematics
Feldkirchenstraße 21
D – 96052 Bamberg, Germany
E-mail: anne.leucht@uni-bamberg.de

Michael H. Neumann
Friedrich-Schiller-Universität Jena
Institut für Mathematik
Ernst-Abbe-Platz 2
D – 07743 Jena
Germany
E-mail: michael.neumann@uni-jena.de

Abstract

We consider a general monotone regression estimation where we allow for independent and dependent regressors. We propose a modification of the classical isotonic least squares estimator and establish its rate of convergence for the integrated L^1 -loss function. The methodology captures the shape of the data without assuming additivity or a parametric form for the regression function. Furthermore, the degree of smoothing is chosen automatically and no auxiliary tuning is required for the theoretical analysis. Some simulations and two real data illustrations complement the study of the proposed estimator.

Keywords: Isotonic least squares estimation, multivariate isotonic regression, nonparametric estimation, rate of convergence, shape constraints, strong mixing.

Submitted: May 2018; First revision: March 2020.

1. INTRODUCTION

We consider the classical mean regression model

$$Y_t = f(I_t) + \varepsilon_t \quad \text{with} \quad E(\varepsilon_t | I_t) = 0 \quad a.s., \quad t \in \mathbb{Z}, \quad (1.1)$$

where we assume that the regression function $f: D \rightarrow \mathbb{R}$, $D \subseteq \mathbb{R}^d$, is unknown and allow for both independent and dependent observations $((Y_t, I_t')'_t)$ (here and in the sequel, x' denotes the transpose of a vector x).

Notably, the problem of estimating a regression function subject to shape constraints, in the context of time series, has not been addressed adequately in the literature, to the best of our knowledge. There exist a large body of literature on estimation and testing for situations where the class of admissible functions f can be parametrized by a finite-dimensional parameter; see e.g. Escanciano (2006), Francq and Zakoian (2010) and Shumway and Stoffer (2011) among others. There are also many results on nonparametric kernel estimators for f relying on the assumption that the covariate vector I_t has a Lebesgue density. For an overview, we refer the reader to the monographs by Härdle (1990) and Fan and Gijbels (1996). On the other hand, there are numerous applications that the covariates do not possess a density with respect to the Lebesgue measure; a case in point is various count time series models which have been employed for the analysis of financial data (e.g. modeling the number of transactions) or biomedical data (e.g. modeling infectious diseases); see Fokianos *et al.* (2009) for instance and Sec. 4.2.

The primary aim of this work is to provide integrated L^1 -loss convergence rate of a nonparametric estimator of f *subject to shape constraints without assuming additivity*; in particular we assume throughout this work that the function f in (1.1) is isotonic. The assumption of isotonicity seems to be appropriate in the context of many applications and, in fact, some popular parametric models share this property, for example, autoregressive and GARCH type models with nonnegative coefficients. Application of standard nonparametric methods such as kernel estimators of the

function f , proposed e.g. by Mukarjee and Stern (1994), Dette *et al.* (2006), Chernozhukov *et al.* (2009), Daouia and Park (2013) or generalized additive modeling studied by Chen and Samworth (2016) among other references, depends on a data-driven choice of smoothing parameters, such as a bandwidth. While the simple leave-one-out cross-validation may fail, the method of leave- k -out cross-validation involves a choice of k , which in turn requires a difficult subjective decision.

Another popular shape-constrained estimator of the function f is the isotonic least squares estimator (LSE) \tilde{f}_n which is given by

$$\tilde{f}_n \in \arg \min_{g \text{ isotonic}} \sum_{t=1}^n (Y_t - g(I_t))^2.$$

In sharp contrast to usual kernel estimators, the isotonic least squares estimator does not require the choice of any smoothing parameter since an appropriate tuning of the degree of smoothing is done automatically. This estimator seems to be less sensitive to irregularities in the design and if the target function is indeed isotonic then this estimator is consistent; see e.g. Christopeit and Tosstorff (1987) and references therein.

Denote by $\mathbb{1}(\cdot)$ the indicator function. It is well known that \tilde{f}_n satisfies at all observation points $x \in \{I_1, \dots, I_n\}$ the following equations:

$$\tilde{f}_n(x) = \max_{U: x \in U} \min_{L: x \in L} \text{Av}_Y(L \cap U) \tag{1.2a}$$

$$= \min_{L: x \in L} \max_{U: x \in U} \text{Av}_Y(L \cap U), \tag{1.2b}$$

where

$$\text{Av}_Y(B) = \frac{\sum_{t=1}^n Y_t \mathbb{1}(I_t \in B)}{\#\{t \leq n: I_t \in B\}}, \quad B \subseteq \mathbb{R}^d,$$

and U and L denote upper and lower sets, respectively; see e.g. Theorem 1 in Brunk (1955) and Theorem 1.4.4 in Robertson *et al.* (1988, p. 23). (A set $U \subseteq \mathbb{R}^d$ is called an upper set if $x \in U$ and $x \leq y$ imply that $y \in U$. Analogously, $L \subseteq \mathbb{R}^d$ is called a lower set if $x \in L$ and $x \geq y$ imply that $y \in L$. Here, the notation $x \leq y$ ($x \geq y$, respectively) denotes that $x_i \leq y_i$ ($x_i \geq y_i$, respectively), for all $i = 1, \dots, d$.) While \tilde{f}_n is uniquely

defined at the observation points, there is some arbitrariness of choosing \tilde{f}_n between these points; only the postulated isotonicity has to be satisfied.

For the univariate case, i.e. $d = 1$, there are already several results reported in the literature concerning the asymptotic behavior (usually assuming a deterministic regressor) of the classical isotonic least squares estimator \tilde{f}_n . Pointwise asymptotic distributions of isotonic least squares estimators assuming short and long range dependence of the error sequence $(\varepsilon_t)_t$ have been derived by Anevski and Hössjer (2006) and Dedecker *et al.* (2011). In particular, it is known that this estimator converges at the optimal rate $n^{-1/3}$ to f . Zhang (2002, Theorem 2.3) studies the case of independent but not necessarily identically distributed errors and shows that $(n^{-1} \sum_{i=1}^n E(\tilde{f}_n(t_i) - f(t_i))^p)^{1/p} = O(n^{-1/3})$, where t_1, \dots, t_n are values of a deterministic covariate and $1 \leq p \leq 3$; see also Chatterjee *et al.* (2015) for a refinement in the case that $p = 2$ but under the assumption of independent and identically distributed errors. Furthermore, Durot (2002, Theorem 1) proves that $E[\int_0^1 |\tilde{f}_n(x) - f(x)| dx] = O(n^{-1/3})$.

However, much less was known about the asymptotic behavior of \tilde{f}_n in the case of multivariate regression models. Hanson *et al.* (1973, Theorem 5) prove uniform consistency of \tilde{f}_n in the case $d = 2$ under the assumptions of deterministic regressors and a continuous target function f . Additionally, these authors provide intuition for the convergence of large deviation probabilities between the estimator and the true regression function towards zero; see Hanson *et al.* (1973, Eq. (26)). Robertson and Wright (1975, Theorems 2.1 and 2.2) state pointwise consistency for \tilde{f}_n in the context of a general partial order for the regressors. Finally, Christopheit and Tosstorff (1987, Theorem 1) prove a consistency result in the d -dimensional case. The authors assume that the errors form a martingale difference sequence and the covariates are continuous and stochastic. Only recently, the empirical L^2 -risk $n^{-1} \sum_{i=1}^n E(\tilde{f}_n(X_i) - f(X_i))^2$ has been investigated for i.i.d. data by Han *et al.* (2019), see also references therein. They derive optimal minimax rates up to a poly-logarithmic factor for lattice designs and discuss random designs, too. In the

latter context, they do not claim optimality of the rates obtained. In particular, they point out that the entropy of the class of isotonic functions in high-dimensions is very large. This results in an enormous amount of possible lower and upper sets involved in computing (1.2a) and (1.2b); see also Gao and Wellner (2007) as well as the discussion in Section 3 in Wu *et al.* (2015).

To the best of our knowledge, there are no results concerning the integrated L^1 convergence rate of isotonic LSE in the case of multivariate regression models available in the literature. Our goal is to fill this gap by proposing a suitable modification of isotonic LSE, so-called block estimators, as described in Section 2. For the case of univariate regression we let intact the isotonic LSE \tilde{f}_n . However, in the multivariate case we propose a slightly simpler estimator by restricting attention to lower and upper sets of (hyper-)rectangular type. As we will show, for both cases of independent regressors (see Theorem 2.1) and dependent data (see Theorem 3.1), such modification avoids the entropy problem and allows derivation of the desired convergence rate. Parallel to our work, this new type of estimator has been investigated by Deng and Zhang (2020) and Han and Zhang (2019) for the case of independent regressors. While the first paper provides rates of convergence for the empirical L^q -risk, the latter derives the limit distribution for the block estimator, after suitable normalization. In sharp contrast to usual nonparametric estimators and in accordance with the classical isotonic LSE, this estimator does not require the choice of an appropriate bandwidth. Such choice could cause problems in the general setting we consider which takes into account possibly irregular distribution of the explanatory variables and dependence among observations. In addition, we allow for an inclusion of a deterministic trend component. Such a covariate accommodates the case of gradual changes over time in contrast to change-point models with stationarity between these points of (abrupt) changes.

The paper is structured as follows. We introduce the proposed estimators and present results on their rate of convergence in Sections 2 (independence case without

trend component) and 3 (dependence case allowing for a deterministic trend). Numerical examples are discussed in Section 4. All proofs as well as technical auxiliary results are deferred to Section 5.

2. MULTIVARIATE ISOTONIC REGRESSION UNDER INDEPENDENCE

Recall (1.1) where we now assume that $f: [0, 1]^d \rightarrow \mathbb{R}$ and $(I'_1, \varepsilon_1)', \dots, (I'_n, \varepsilon_n)'$ are independent and identically distributed random variables on a probability space (Ω, \mathcal{A}, P) . We assume that the conditional mean function f is isotonic, that is, monotonically non-decreasing in each argument. Following the discussion of Section 1, we consider estimators defined at the design points x by

$$\widehat{f}_n^{(max-min)}(x) = \max_{a: a \leq x} \min_{b: b \geq x} \text{Av}_Y([a, b]) \quad (2.1a)$$

and

$$\widehat{f}_n^{(min-max)}(x) = \min_{b: b \geq x} \max_{a: a \leq x} \text{Av}_Y([a, b]), \quad (2.1b)$$

where, for $a, b \in [0, 1]^d$, $[a, b] = [a_1, b_1] \times \dots \times [a_d, b_d]$. As pointed out by Deng and Zhang (2020), (2.1a) and (2.1b) have to be modified for x not being a design point. Since it could well happen that a rectangle with $x \in [a, b]$ does not contain any design point we set $n_{a,b} = \#\{t \leq n: I_t \in [a, b]\}$, $n_{a,*} = \#\{t \leq n: a \leq I_t\}$, $n_{*,b} = \#\{t \leq n: I_t \leq b\}$ and define

$$\widehat{f}_n^{(max-min)}(x) = \max_{a: a \leq x, n_{a,*} > 0} \min_{b: b \geq x, n_{a,b} > 0} \text{Av}_Y([a, b]) \quad (2.2a)$$

and

$$\widehat{f}_n^{(min-max)}(x) = \min_{b: b \geq x, n_{*,b} > 0} \max_{a: a \leq x, n_{a,b} > 0} \text{Av}_Y([a, b]), \quad (2.2b)$$

It follows from the construction of both $\widehat{f}_n^{(max-min)}(x)$ and $\widehat{f}_n^{(min-max)}(x)$ that they are isotonic. We define the isotonic estimator \widehat{f}_n of f by

$$\widehat{f}_n(x) = \left(\widehat{f}_n^{(max-min)}(x) + \widehat{f}_n^{(min-max)}(x) \right) / 2. \quad (2.3)$$

In the univariate case, \widehat{f}_n is equal to \widetilde{f}_n at the observation points. The proposed estimator deviates from \widetilde{f}_n in the multivariate case though. Also note that $\widehat{f}_n^{(max-min)}(x) \leq \widehat{f}_n^{(min-max)}(x)$ if x is a design point. However, this does in general

not hold true if x is not a design point; see Deng and Zhang (2020, Section 2) for an example. The proofs of Theorems 2.1 and 3.1 below show that replacing lower and upper sets by hyperrectangles in (1.2a) and (1.2b) simplifies the derivation of the desired rate of convergence and its computation.

Firstly, we study the case of independent and identically distributed variables $(I_t, \varepsilon_t)'$. We impose the following condition.

(A1) (i) The information variables I_t possess a density p on $[0, 1]^d$, such that

$$C_1 := \inf_{x \in [0, 1]^d} p(x) \leq \sup_{x \in [0, 1]^d} p(x) =: C_2,$$

where $0 < C_1 \leq C_2 < \infty$.

(ii) The error sequence $(\varepsilon_t)_{t \in \mathbb{N}}$ satisfies

$$E(\varepsilon_t | I_t) = 0 \text{ a.s.}, \quad \text{and} \quad E(\varepsilon_t^2 | I_t) \leq \bar{\sigma}_\varepsilon^2 \text{ a.s.},$$

where $\bar{\sigma}_\varepsilon^2 < \infty$.

It is well known that the traditional isotonic estimator $\tilde{f}_n(x)$ is problematic when x is close to the boundary of the support of the distribution of the I_t ; see e.g. the discussion in Sampson *et al.* (2003). The same is true for \hat{f}_n at points x near the boundary of the domain. To fix the bias problem at extreme small and large design points, Wu *et al.* (2015) proposed an adequate modification by pulling up and down the isotonic LSE at these particular locations. This, however, requires some sort of tuning parameter whose appropriate choice is somewhat subjective. We show that our estimator \widehat{f}_n achieves the optimal rate of convergence if we neglect its behavior near the boundary and focus on estimating f on the set

$$D_n = (1/M_n, 1 - 1/M_n]^d \quad \text{with} \quad M_n = \lceil n^{1/(d+2)} \rceil. \quad (2.4)$$

Note that $n^{-1/(d+2)}$ corresponds to an asymptotically mean square error-optimal bandwidth when the function to be estimated has a degree of smoothness 1. If

lower and upper bounds \underline{f} and \overline{f} for f are known, then we could avoid unsatisfactory behavior of \widehat{f}_n near the boundary or if the (random) design is too irregular by setting

$$\widehat{\widehat{f}}_n(x) = \min \{ \max \{ \widehat{f}_n(x), \underline{f} \}, \overline{f} \}.$$

Under minimal assumptions and assuming existence of second moments for the error terms we prove the following theorem.

Theorem 2.1. *Suppose that Assumption (A1) holds true. Then, with λ^d denoting the Lebesgue measure on \mathbb{R}^d ,*

- (i) $\int_{D_n} |\widehat{f}_n(x) - f(x)| \lambda^d(dx) = O_P(n^{-1/(d+2)}).$
- (ii) *If in addition $\underline{f} \leq f(x) \leq \overline{f}$ for all $x \in [0, 1]^d$, then*

$$E \left[\int_{[0,1]^d} |\widehat{\widehat{f}}_n(x) - f(x)| \lambda^d(dx) \right] = O(n^{-1/(d+2)}).$$

Consider the special case of a partially differentiable function $f: [0, 1]^d \rightarrow \mathbb{R}$. Then the assumption of isotonicity implies that

$$\int_{[0,1]^d} \sum_{i=1}^d |\partial_i f(x)| \lambda^d(dx) \leq d(\sup_x \{f(x)\} - \inf_x \{f(x)\}).$$

Hence, the degree of smoothness, say β , measured in the L^1 -norm, is equal to 1. It is well known that, under appropriate conditions, the optimal rate of convergence for the L^1 -loss is $n^{-\beta/(2\beta+d)}$ which reduces to $n^{-1/(d+2)}$, when $\beta = 1$; see Stone (1982). Hence, Theorem 2.1 indicates that \widehat{f}_n achieves the optimal rate of convergence in the class of isotonic functions. Recall that in contrast to the classical isotonic LSE which is obtained by using all possible lower and upper sets in (1.2a) and (1.2b) our estimator \widehat{f}_n is based on averages over hyperrectangles only. This reduced complexity allows us to derive the desired rate of convergence.

3. MULTIVARIATE ISOTONIC REGRESSION UNDER DEPENDENCE

Recall again (1.1) where we now allow the random variables to be dependent. We assume the information variables to be of the form $I_{n,t} = (X'_{n,t}, Z'_{n,t})'$, where $X_{n,t}$ is a d_1 -dimensional vector consisting of components with values in $\mathbb{N}_0 = \{0, 1, \dots\}$, and $Z_{n,t}$ is a d_2 -dimensional covariate consisting of variables with continuous marginal distribution functions and possibly a deterministic trend component t/n . Here, we allow for $d_1, d_2 \in \mathbb{N}_0$ with $d = d_1 + d_2 > 0$. Note that by setting $d_1 = 0$ or $d_2 = 0$, it is possible that $I_{n,t}$ is just equal to $X_{n,t}$ or $Z_{n,t}$, respectively. More specifically, we distinguish between two cases: either the covariate vector $Z_{n,t}$ includes a trend component of the form t/n , i.e. $Z_t = Z_{n,t} = (\tilde{Z}'_t, t/n)'$, where \tilde{Z}_t denotes the rest of the covariates, or the covariate vector is free of a trend. In this section, we consider again the isotonic estimator \hat{f}_n defined by (2.3). We show that the results of Theorem 2.1 can be generalized to the case of strong mixing random variables provided that we impose some additional assumptions. We suppose that:

(A2) (i) The error sequence $(\varepsilon_{n,t})_{t \in \mathbb{N}}$ satisfies

$$E(\varepsilon_{n,t} \mid I_{n,1}, \dots, I_{n,t}, \varepsilon_{n,1}, \dots, \varepsilon_{n,t-1}) = 0 \quad \text{a.s.},$$

$$E(\varepsilon_{n,t}^2 \mid I_{n,1}, \dots, I_{n,t}, \varepsilon_{n,1}, \dots, \varepsilon_{n,t-1}) \leq \bar{\sigma}_\varepsilon^2 \quad \text{a.s.},$$

where $\bar{\sigma}_\varepsilon^2 < \infty$.

(ii) The process $(I_{n,t})_{t \in \mathbb{N}}$ is strong (α) -mixing with corresponding mixing coefficients satisfying

$$\alpha(r) = O(r^{-d_2-1/2}).$$

(iii) The function $f: D \rightarrow \mathbb{R}$, $D \subseteq \mathbb{R}^d$, is bounded.

Having in mind that $I_{n,t} = (X'_{n,t}, Z'_{n,t})'$ contains $d_1 \geq 0$ components with a discrete distribution and $d_2 \geq 0$ components having either a continuous distribution or being nonrandom such as t/n we impose the following condition:

(A3) For $t = 1, \dots, n$ and $n \in \mathbb{N}$, the random vectors $Z_{n,t}$ consist of components with continuous marginal distribution functions and/or a deterministic trend component t/n .

- (i) There exist continuous distribution functions G_1, \dots, G_{d_2} on \mathbb{R} and, for all $K \in \mathbb{N}$, constants $0 < C_1 = C_1(K) \leq C_2 = C_2(K) < \infty$ such that $\forall k_1, \dots, k_{d_1} \leq K, \forall a_i \leq b_i$

$$\begin{aligned} C_1 \prod_{i=1}^{d_2} (G_i(b_i) - G_i(a_i)) & - \frac{1}{n} \\ & \leq \frac{1}{n} \sum_{t=1}^n P(X_{n,t} = (k_1, \dots, k_{d_1})', Z_{n,t} \in (a_1, b_1] \times \dots \times (a_{d_2}, b_{d_2}]) \\ & \leq \frac{1}{n} \sum_{t=1}^n P(Z_{n,t} \in (a_1, b_1] \times \dots \times (a_{d_2}, b_{d_2}]) \\ & \leq C_2 \prod_{i=1}^{d_2} (G_i(b_i) - G_i(a_i)) + \frac{1}{n}. \end{aligned}$$

- (ii) There exists some constant $C_3 < \infty$ such that, for all d -dimensional hyperrectangles C ,

$$P\left(I_{n,t} \in C \mid I_{n,1}, \dots, I_{n,t-d}, \varepsilon_{n,1}, \dots, \varepsilon_{n,t-d}\right) \leq C_3 P(I_{n,t} \in C).$$

Before we proceed, some comments on assumption (A3) are in order. Condition (A3)(i) means that the ‘‘average distribution’’ of the continuous random variables behaves as a d_2 -dimensional product distribution which has, after an appropriate rescaling with $G_1^{-1}, \dots, G_{d_2}^{-1}$, a density bounded away from zero on $[0, 1]^{d_2}$. The terms $\pm 1/n$ are needed to accommodate the possible case of a deterministic trend variable t/n . Also note that assumption (A1)(i) implies the validity of assumption (A3)(ii). We impose a condition on $P(I_{n,t} \in C \mid I_{n,1}, \dots, I_{n,t-d}, \varepsilon_{n,1}, \dots, \varepsilon_{n,t-d})$ rather than $P(I_{n,t} \in C \mid I_{n,1}, \dots, I_{n,t-1}, \varepsilon_{n,1}, \dots, \varepsilon_{n,t-1})$ in order to accommodate the case where $I_{n,t} = (Y_{n,t-1}, \dots, Y_{n,t-d})'$. Scaling the deterministic trend by n is a common approach in the literature on non-stationary time series; see e.g. Dahlhaus (1997) and Dahlhaus and Neumann (2001). Under suitable regularity conditions, it allows

for the estimation of the influence of the rescaled time component on the regression function (rescaled to $[0,1]$) via so-called infill asymptotics.

To simplify the notation, we suppress the index n in $Y_{n,t}$, $I_{n,t}$ and $\varepsilon_{n,t}$ from here on, just keeping in mind that also a triangular scheme is allowed, e.g., when a trend variable t/n is included.

Recall again that the traditional isotonic estimator $\tilde{f}_n(x)$ is problematic when x is close to the boundary of the support of the distribution of the I_t . We neglect the behavior of \hat{f}_n near the boundary and focus primarily on estimating f on

$$\tilde{D}_n = \{0, \dots, K\}^{d_1} \times (G_1^{-1}(\tilde{h}_n), G_1^{-1}(1 - \tilde{h}_n)] \times \dots \times (G_{d_2}^{-1}(\tilde{h}_n), G_{d_2}^{-1}(1 - \tilde{h}_n)],$$

where $\tilde{h}_n = 1/\tilde{M}_n$ and $\tilde{M}_n = \lceil n^{1/(d_2+2)} \rceil$. As in Section 2, if lower and upper bounds \underline{f} and \bar{f} for $f(x)$ are known, we can take this into account by setting

$$\widehat{f}_n(x) = \min \{ \max \{ \hat{f}_n(x), \underline{f} \}, \bar{f} \}.$$

In this case, we evaluate our estimator \widehat{f}_n on the set $\tilde{D} = \{0, \dots, K\}^{d_1} \times [0, 1]^{d_2}$. Denote by Q_1, \dots, Q_{d_2} the probability measures corresponding to the distribution functions G_1, \dots, G_{d_2} , respectively. With μ^{d_1} being the counting measure on $\mathbb{N}_0^{d_1}$, define $\nu = \mu^{d_1} \otimes Q_1 \otimes \dots \otimes Q_{d_2}$.

Theorem 3.1. *Suppose that Assumptions (A2) and (A3) hold true. Then,*

- (i) $\int_{\tilde{D}_n} |\hat{f}_n(x) - f(x)| \nu(dx) = O_P(n^{-1/(d_2+2)}).$
- (ii) *If in addition $\underline{f} \leq f(x) \leq \bar{f}$ for all $x \in \tilde{D}$, then*

$$E \left[\int_{\tilde{D}} |\widehat{f}_n(x) - f(x)| \nu(dx) \right] = O(n^{-1/(d_2+2)}).$$

Some remarks are in order. We point out that the obtained rate of convergence does not depend on the number of discrete explanatory random variables. This is explained by the fact that, for any $k_1, \dots, k_{d_1} \in \{0, \dots, K\}$, the cardinality of the set $\{t \leq n: X_{n,t} = (k_1, \dots, k_{d_1})'\}$ is proportional to the sample size n . Therefore, there

is no need to smooth over the first d_1 directions and there is no loss due to a trade-off between bias and variance that would appear with nonparametric smoothing techniques.

Properties of the noise process can be taken into account, provided that we have some prior knowledge. Indeed, if we knew the conditional variance $E(\varepsilon_t^2 | I_t)$, e.g. in the case of a known distributional family for the errors, then we could replace the means $\text{Av}_Y(B) = \sum_{t=1}^n Y_t \mathbb{1}(I_t \in B) / \#\{t \leq n: I_t \in B\}$ by the weighted means $\sum_{t=1}^n w(I_t) Y_t \mathbb{1}(I_t \in B) / \sum_{t=1}^n w(I_t) \mathbb{1}(I_t \in B)$, where the weights $w(I_t)$ are proportional to $1/E(\varepsilon_t^2 | I_t)$. This corresponds to a weighted least squares estimator in linear regression. However, our main intention was to produce a general, fully non-parametric method. Since prior knowledge of $E(\varepsilon_t^2 | I_t)$ is rarely available, we pursue the approach based on unweighted means.

Example 3.1. Suppose that integer-valued random variables Y_0, Y_1, \dots, Y_n are observed, where

$$Y_t | \mathcal{F}_{t-1} \sim \text{Poisson}(\lambda_t) \quad \lambda_t = f(Y_{t-1}, Z_{t-1}),$$

and Z_t is a covariate with values in $[0, 1]^{d_2}$ which is independent of $Y_t, \dots, Y_0, Z_{t-1}, \dots, Z_0$, $\mathcal{F}_s = \sigma(Y_s, Z_s, \dots, Y_0, Z_0)$. Assume that the function $f: \mathbb{N}_0 \times [0, 1]^{d_2} \rightarrow [0, M]$ is isotonic and bounded by $M < \infty$. The information variable at time t is $I_t = (Y_{t-1}, Z'_{t-1})'$. We have that

$$Y_t = f(I_t) + \varepsilon_t,$$

where

$$E(\varepsilon_t | I_t, \dots, I_0, \varepsilon_{t-1}, \dots, \varepsilon_0) = 0,$$

$$E(\varepsilon_t^2 | I_t, \dots, I_0, \varepsilon_{t-1}, \dots, \varepsilon_0) = f(I_t) \leq M.$$

It can be shown that Assumption (A2)(i) is also fulfilled. Indeed, let $Q_t^k := P^{Y_t | Y_{t-1}=k} = \int \text{Poisson}(f(k, z)) P^{Z_{t-1}}(dz)$. Since $f(k, z) \in [0, M]$ for all values of k and z ,

$$\inf_t \inf_{k \in \mathbb{N}_0} Q_t^k(\{0\}) > 0,$$

that is, Doeblin's condition is fulfilled. It follows from Theorem 2.4.1 on page 88 in Doukhan (1994) that the Markov chain $(Y_t)_t$ is uniformly (ϕ -) mixing and, therefore, absolutely regular with coefficients satisfying

$$\beta^Y(k) \leq C\rho^k \quad \forall k \in \mathbb{N}_0,$$

for some $C < \infty$ and $\rho \in [0, 1)$. Since the process $(I_t)_t$ is also a Markov chain, we obtain that

$$\begin{aligned} & \beta(\sigma(I_0, I_1, \dots, I_t), \sigma(I_{t+k}, I_{t+k+1}, \dots)) \\ &= \beta(\sigma(I_t), \sigma(I_{t+k})) \\ &\leq \beta(\sigma(Y_t, I_t), \sigma(I_{t+k})) \\ &= \beta(\sigma(Y_t), \sigma(I_{t+k})) \\ &= \beta(\sigma(Y_t), \sigma(Y_{t+k-1})). \end{aligned}$$

(The first and the last but one equalities follows from the Markovian structure; see also the note after Theorem 7.3 in Bradley (2007). The last one follows from independence of Z_{t+k-1} and (Y_t, Y_{t+k-1}) ; see also Theorem 6.2 in Bradley (2007).) Hence, the coefficients of absolute regularity of the process $(I_t)_t$ satisfy

$$\beta^I(k) \leq \beta^Y(k-1) \leq C\rho^{k-1} \quad \forall k \in \mathbb{N}.$$

4. APPLICATIONS

4.1. Simulations. We illustrate the theoretical results by a simulation study comparing the performance of \widehat{f}_n and the isotonic LSE \widetilde{f}_n in terms of their average biases and L^1 , L^2 errors over a specified grid. Hence, if x belongs to a rectangular grid G with B points, we evaluate the average bias by $\sum_{x \in G} (\widehat{f}_n(x) - f(x))/B$, the L^1 error by $\sum_{x \in G} |\widehat{f}_n(x) - f(x)|/B$ and the L^2 error by $\sum_{x \in G} (\widehat{f}_n(x) - f(x))^2/B$ for \widehat{f}_n , and similarly for \widetilde{f}_n .

We study an additive and a non-additive model under independence and dependence. Given some parameter values a, b, c, d such that $a, b, c, d > 0$, consider the following isotonic functions

$$f(i_1, i_2) = (d + 1) + ai_1(1 - \exp(-bi_2)) \quad (4.1a)$$

and

$$f(i_1, i_2) = d + \frac{a}{1 + \exp(-ci_1)} + bi_2. \quad (4.1b)$$

In the case of independent data and for the function (4.1a) we generate data as $Y_t = f(I_{t1}, I_{t2}) + \varepsilon_t$, $t = 1, 2, \dots, n$, where the covariates I_{t1} and I_{t2} are independent $U(0, 1)$ random variables and the error process $(\varepsilon_t)_t$ consists of iid centered exponential random variables. Similarly, for the model given by (4.1b) we generate independent Poisson distributed random variables with mean $\lambda_t = f(I_{t1}, I_{t2})$. In the time series cases, set $I_t = (Y_{t-1}, Z_t)'$ where $Z_t = t/n$ and generate data $Y_t = f(Y_{t-1}, Z_t) + \varepsilon_t$ using (4.1a) by assuming identical error structure and $a < 1$. Similarly for the Poisson model we use $I_t = (Y_{t-1}, Z_t)'$ and generate data as in the Example 3.1.

We compute the isotonic least squares estimator by using the R package `isotonic.pen` which returns the values of the estimated function on an equidistant 21×21 grid; see Wu *et al.* (2015) for details. Note that the estimator \widehat{f}_n , defined by (2.3), can be computed exactly at a given point by a direct approach which requires at most $O(n^2)$ calculations, regardless the dimension d . The total computation time for n points is $O(n^3)$, see also Han and Zhang (2019, page 4). Since the goal of this study is to compare the risks of the estimators we refrain discussing any computational algorithms. To compare \widehat{f}_n with the isotonic LSE we use the grid computed by `isotonic.pen` discarding some points so that we bypass possible boundary issues. To this end, we choose the lower left / upper right corner of the grid such that, on the one hand, the number of observed information variables within the corresponding rectangle is large and on the other hand, for every grid point there is at least one

upper and one lower set containing data points. To compare the empirical performance of the estimators, we compute the integrated L^1 , L^2 errors and the average biases, over the grid values, as defined before. This process is repeated 500 times.

Table 1 reports results of this study for the case of the non-additive model (4.1a) and for a specific choice of parameter values. It is seen that \widehat{f}_n outperforms the isotonic LSE in terms of L^1 and L^2 errors for all sample sizes considered. Both estimators have positive bias in the case of independent data but this changes for time series data. Figure 1 shows box plots of the values of integrated L^1 error for three different sample sizes and for different parameter values. These results, which are obtained by considering time series data, reinforce the previous findings. Similar results are obtained for the case of the additive model (4.1b) under a conditional Poisson model, see Table 2 and Figure 2.

	\widehat{f}_n			\widetilde{f}_n		
Sample Size	Av. Bias	L^1	L^2	Av. Bias	L^1	L^2
Independent Data						
250	0.8393	0.8393	0.7212	0.8856	0.8857	0.8524
500	0.5883	0.5887	0.3705	0.8801	0.8802	0.8256
1000	0.2871	0.2897	0.1041	0.8726	0.8727	0.8167
Time Series Data						
250	-0.3277	0.3626	0.2180	0.7396	0.7513	0.7209
500	-0.2827	0.3623	0.2236	0.6874	0.7064	0.6500
1000	-0.2349	0.3625	0.2136	0.6326	0.6579	0.5781

TABLE 1. Average bias, L^1 and L^2 error of \widehat{f}_n and \widetilde{f}_n for the non-additive model (4.1a) for $d = 0.2$, $a = 0.1$ and $b = 0.7$ and for different sample sizes under independence and dependence. Data are generated with centered exponential innovations.

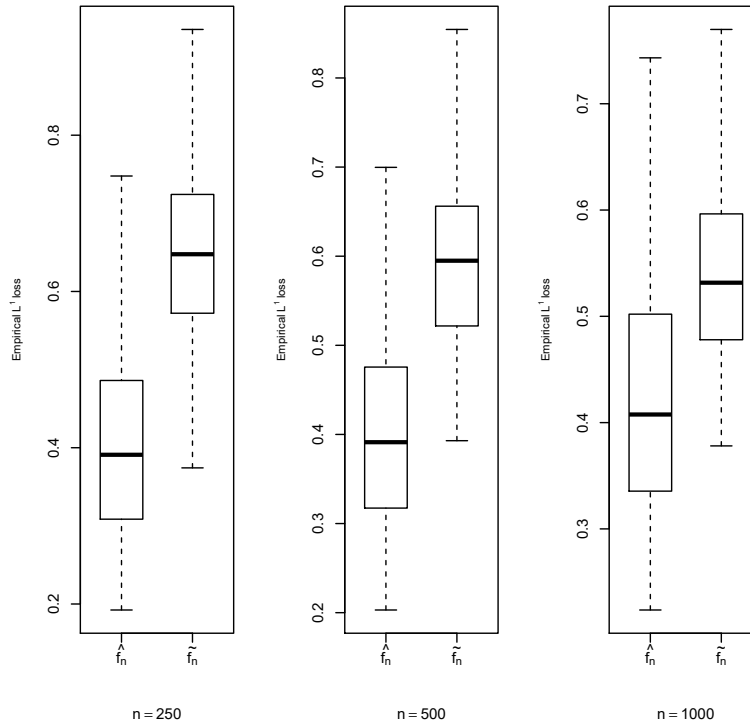


FIGURE 1. Box plots of empirical L^1 loss values for \hat{f}_n and the isotonic LSE \tilde{f}_n for the non-additive model (4.1a) for $d = 0.2$, $a = 0.1$ and $b = 0.2$ under dependence. Data are generated with centered exponential innovations.

4.2. Data Examples. We apply the methodology to biological and financial time series which exhibit non-stationarity to illustrate the applicability of isotonic estimation. First, we investigate the population growth of whooping cranes that became nearly extinct during the period 1938-1955. Whooping cranes are one of the largest birds in North America but also one of the rarest that can be found in the continent. For some time their population has been constantly decreasing and reached to about 20 individual birds in the world. With the employment of various conservation measures the population grew over the last years. The data we have are depicted in Figure 3 which shows the growth of population of whooping cranes between 1938 to 2005; see Int. Recovery Plan (2007). Note that this is a case of an integer valued

	\widehat{f}_n			\widetilde{f}_n		
Sample Size	Av. Bias	L^1	L^2	Av. Bias	L^1	L^2
Independent Data						
250	-0.8999	0.9324	1.0219	0.5771	0.7790	0.9063
500	0.4590	0.5416	0.4575	0.5646	0.7473	0.8017
1000	-0.1573	0.4181	0.2947	0.5633	0.7313	0.7436
Time Series Data						
250	-1.0037	1.0207	1.2310	1.0912	1.1292	1.6985
500	-0.6787	0.7823	0.7915	1.0803	1.1218	1.6747
1000	-0.4199	0.6647	0.5707	1.0620	1.1146	1.6664

TABLE 2. Average bias, L^1 and L^2 error of \widehat{f}_n and \widetilde{f}_n for the additive model (4.1b) for $d = 2$, $a = 0.5$, $c = 0.3$ and $b = 1.7$ and for different sample sizes under independence and dependence. Data are generated according to a Poisson model.

time series. The second example refers to daily net asset value (NAV) of the Black-Rock Global Allocation Fund during the period 1/4/2016 to 30/1/2018. Here we note that the series takes values on real numbers.

For both of these data examples, a simple time series plot reveals increasing trend and strong autocorrelation which decays slowly. The partial autocorrelation functions shows a strong autocorrelation at lag 1; see the upper panel of Figures 3 and 4. We fit a non-parametric time series model to these data by using isotonic estimation methods. We include the covariate vector $I_t = (Y_{t-1}, t/n)'$, where n is the number of effective observations (e.g. for the population growth of whooping cranes the number of observation is equal to 68 but $n = 67$ because of the inclusion of Y_{t-1}). We consider again the estimator \widehat{f}_n and the isotonic LSE \widetilde{f}_n and work the same way as it was explained in Subsection 4.1. The lower panels of Figures 3 and 4 show that both estimators are quite close near the observation points, but differ significantly at some grid points that are located far from the bulk of data. We

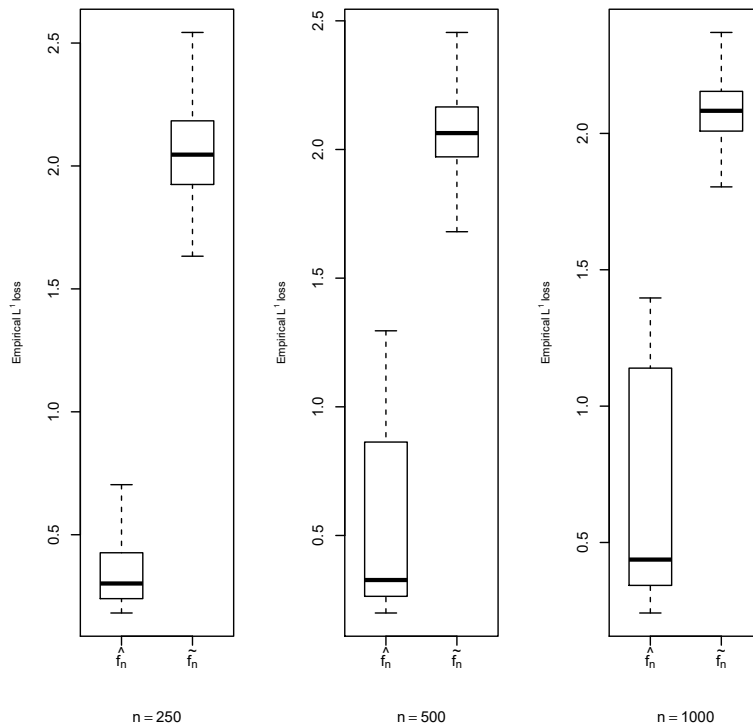


FIGURE 2. Box plots of empirical L^1 loss values for \widehat{f}_n and the isotonic LSE \widetilde{f}_n for the additive model (4.1b) for $d = 0.5$, $a = 1.5$, $c = 0.6$ and $b = 2$ under dependence. Data are generated according to a Poisson model.

examine the performance of both methods for estimating the two models. This task is accomplished by studying the in sample predictive power using the mean absolute prediction error (MAPE), that is $\sum_{t=1}^n |\widehat{Y}_t - Y_t|/n$ and the mean square error (MSE) given by $\sum_{t=1}^n (\widehat{Y}_t - Y_t)^2/n$. Here, \widehat{Y}_t is obtained by evaluating \widehat{f}_n and \widetilde{f}_n , respectively, on a grid point close to $(Y_{t-1}, t/n)'$. The results are shown in Table 3. Clearly, the new estimator \widehat{f}_n performs better in terms of MAPE, but not in terms of MSE.

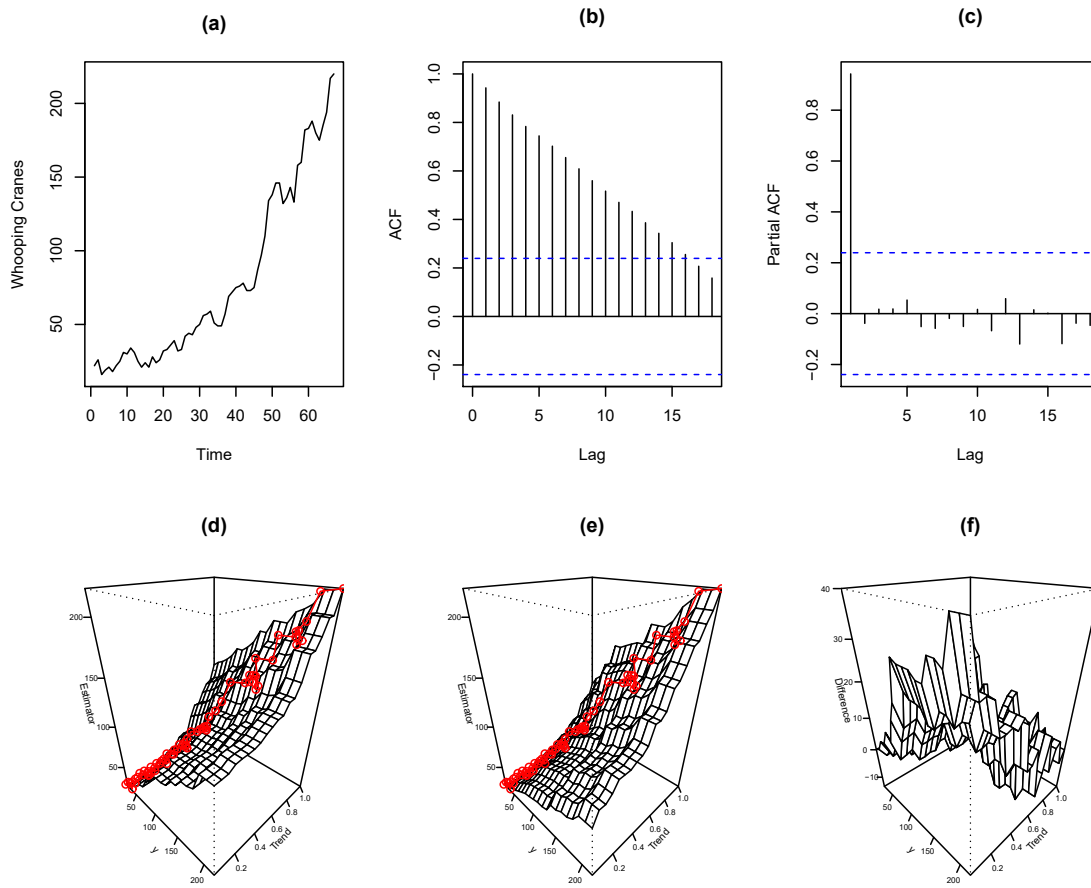


FIGURE 3. (a) Time series plot of the yearly number of whooping cranes between 1938 to 2005. (b) Autocorrelation function. (c) Partial autocorrelation function. (d) Plot of isotonic LSE \tilde{f}_n and the data (red points). (e) Plot of the estimator \hat{f}_n and the data (red points). (f) Plot of the difference $\tilde{f}_n - \hat{f}_n$.

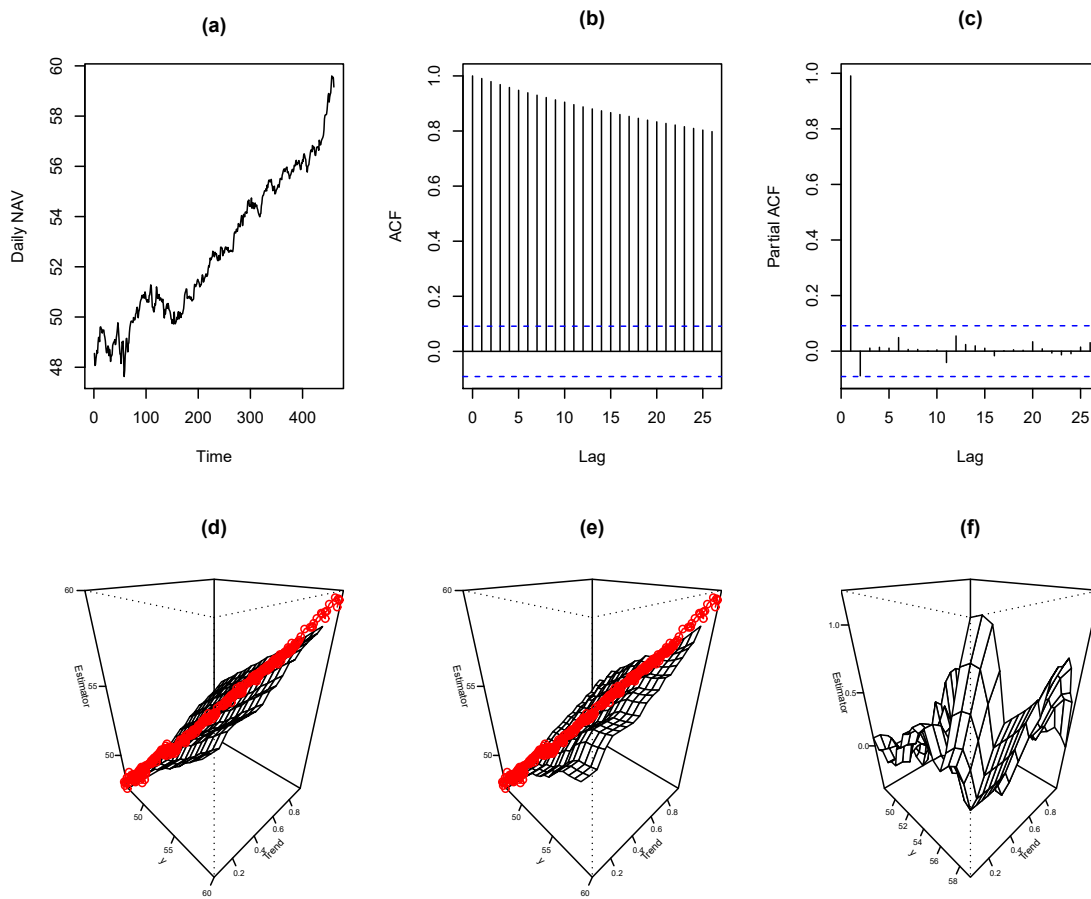


FIGURE 4. (a) Time series plot of daily NAV prices of BlackRock Global Allocation Fund during the period 1/4/2016 to 30/1/2018. (b) Autocorrelation function. (c) Partial autocorrelation function. (d) Plot of isotonic LSE \tilde{f}_n and the data (red points). (e) Plot of the estimator \hat{f}_n and the data (red points). (f) Plot of the difference $\tilde{f}_n - \hat{f}_n$.

	Example 1 (Whooping cranes)		Example 2 (BlackRock Global Allocation Fund)	
	MAPE	MSE	MAPE	MSE
\widehat{f}_n	5.916	65.479	0.311	0.158
\widetilde{f}_n	6.002	64.424	0.312	0.156

TABLE 3. MAPE and MSE of \widehat{f}_n and the isotonic LSE \widetilde{f}_n after fitting the isotonic regression models to real data.

5. PROOFS AND AUXILIARY RESULTS

We prove our main results in Section 5.1. Some auxiliary lemmas are stated and proved in Section 5.2.

5.1. Proofs of the main results.

Proof of Theorem 2.1. The estimator \widehat{f}_n is based on means over hyperrectangles. Note that $h_n = 1/M_n$ (with M_n defined in (2.4)) corresponds to an asymptotically mean square error-optimal bandwidth of a kernel estimator when the function to be estimated has a degree of smoothness 1. Having this in mind, we define, for multi-indexes $k = (k_1, \dots, k_d)$, grid points by

$$x_k = (k_1 h_n, \dots, k_d h_n)', \quad (0 \leq k_i \leq M_n \quad \forall i)$$

and we split $[0, 1]^d$ into subsets

$$B_k = (x_{k-1}, x_k] = ((k_1 - 1)h_n, k_1 h_n] \times \dots \times ((k_d - 1)h_n, k_d h_n] \quad \forall k \in K_n,$$

where $K_n = \{k: 1 \leq k_1, \dots, k_d \leq M_n\}$. We expect a *regular* behavior of \widehat{f}_n if there are sufficiently many observations in each box B_k . Recall that C_1 is the lower bound on the density of the information variables I_t which is assumed to exist by (A1)(i). Then, regularity of \widehat{f}_n is guaranteed to hold, provided that the event

$$A_n = \left\{ \omega: \#\{t \leq n: I_t(\omega) \in B_k\} \geq (C_1/2) n^{2/(d+2)} \quad \forall k \in K_n \right\} \quad (5.1)$$

occurs with probability tending to one. We now prove both assertions of the Theorem.

Proof of (i): In view of assertion (ii), we show a slightly stronger result, namely,

$$P(A_n^c) = O\left(n^{-1/(d+2)}\right). \quad (5.2)$$

Since $\sum_{t=1}^n P(I_t \in B_k) \geq n C_1 h_n^d \geq C_1 n^{2/(d+2)}$ it suffices to show that

$$\sum_{k \in K_n} P\left(\sum_{t=1}^n P(I_t \in B_k) - \mathbb{1}(I_t \in B_k) \geq \frac{C_1 n^{2/(d+2)}}{2}\right) = O\left(n^{-1/(d+2)}\right), \quad (5.3)$$

where $\mathbb{1}(\cdot)$ denotes the indicator function. We obtain from Bernstein's inequality, for all $k \in K_n$ and $\mu_n = C_1 n^{2/(d+2)}/2$, that

$$\begin{aligned} P\left(\sum_{t=1}^n P(I_t \in B_k) - \mathbb{1}(I_t \in B_k) \geq \mu_n\right) &\leq \exp\left\{-\frac{\mu_n^2/2}{n P(I_1 \in B_k) (1 - P(I_1 \in B_k)) + \mu_n/3}\right\} \\ &\leq \exp\{-C n^{2/(d+2)}\}, \end{aligned}$$

for some $C > 0$, which proves (5.3) and therefore (5.2).

We analyze the contribution of the stochastic part and the bias of the estimator separately. For the latter, we exploit the assumed isotonicity in conjunction with boundedness of f in order to construct an estimate of the integrated bias from above and below. To this end, denote for an arbitrary function g its positive (respectively negative) part by g_+ (respectively g_-). Then, it suffices to show that

$$E\left[\int_{D_n} (\widehat{f}_n(x) - f(x))_+ \lambda^d(dx) \mathbb{1}_{A_n}\right] = O(n^{-1/(d+2)}), \quad (5.4a)$$

and

$$E\left[\int_{D_n} (\widehat{f}_n(x) - f(x))_- \lambda^d(dx) \mathbb{1}_{A_n}\right] = O(n^{-1/(d+2)}). \quad (5.4b)$$

Note that D_n introduced in (2.4) satisfies $D_n = \bigcup_{k: 1 < k_1, \dots, k_d < M_n} B_k$. If the event A_n occurs then the set $(x_k, x_{k+1}]$ is non-empty. Therefore, for $x \in B_k = (x_{k-1}, x_k]$,

$$\widehat{f}_n^{(max-min)}(x), \widehat{f}_n^{(min-max)}(x) \leq \sup_{y \leq x_k} \text{Av}_Y((y, x_{k+1}]),$$

which implies that

$$\begin{aligned} (\widehat{f}_n(x) - f(x))_+ \mathbb{1}_{A_n} &\leq \left(\sup_{y \leq x_k} \text{Av}_Y((y, x_{k+1}]) - f(x)\right)_+ \mathbb{1}_{A_n} \\ &\leq \sup_{y \leq x_k} |\text{Av}_\varepsilon((y, x_{k+1}])| \mathbb{1}_{A_n} + (f(x_{k+1}) - f(x)). \end{aligned} \quad (5.5)$$

By Lemma 5.1 and since $\lambda^d(B_k) = h_n^d = O(n^{-d/(d+2)})$ we obtain for the bias that

$$\begin{aligned} &\sum_{k: 1 < k_1, \dots, k_d < M_n} \int_{B_k} (f(x_{k+1}) - f(x)) \lambda^d(dx) \\ &\leq \sum_{k: 1 < k_1, \dots, k_d < M_n} (f(x_{k+1}) - f(x_{k-1})) \lambda^d(B_k) \\ &\leq 2d M_n^{d-1} (f(1, \dots, 1) - f(0, \dots, 0)) h_n^d = O(n^{-1/(d+2)}). \end{aligned} \quad (5.6)$$

For the stochastic part, we estimate $E[\sup_{y \leq x_k} |\text{Av}_\varepsilon((y, x_{k+1}])|]$. For this purpose, we define a dyadic scheme of nested hyperrectangles: For $j_1, \dots, j_d \geq 0$,

$$B_k^{(j_1, \dots, j_d)} = ((k_1 + 1 - 2^{j_1})h_n, (k_1 + 1)h_n] \times \dots \times ((k_d + 1 - 2^{j_d})h_n, (k_d + 1)h_n].$$

(We have in particular $B_k^{(0, \dots, 0)} = B_{k+1}$ and $B_k^{(j_1, \dots, j_d)} = \bigcup_{m_1, \dots, m_d: 0 \leq m_i \leq 2^{j_i-1}} B_{(k_1+1-m_1, \dots, k_d+1-m_d)} \cdot$)

Since $\{y \leq x_k\} \subseteq \bigcup_{j_1, \dots, j_d \geq 0} B_k^{(j_1+1, \dots, j_d+1)} \setminus B_k^{(j_1, \dots, j_d)}$ and since $y \in B_k^{(j_1+1, \dots, j_d+1)} \setminus B_k^{(j_1, \dots, j_d)}$ implies that $(y, x_{k+1}] \supseteq B_k^{(j_1, \dots, j_d)}$ we obtain, for all $x \in B_k$,

$$\sup_{y \leq x_k} |\text{Av}_\varepsilon((y, x_{k+1}])| \mathbb{1}_{A_n} \leq \sum_{j_1, \dots, j_d \geq 0} \frac{\sup \left\{ \left| \sum_{t=1}^n \varepsilon_t \mathbb{1}(I_t \in (y, x_{k+1}]) \right| : y \in B_k^{(j_1+1, \dots, j_d+1)} \setminus B_k^{(j_1, \dots, j_d)} \right\}}{\#\{t \leq n: I_t \in B_k^{(j_1, \dots, j_d)}\}} \mathbb{1}_{A_n}.$$

Recall that if the event A_n occurs, then $\#\{t \leq n: I_t \in B_k\} \geq (C_1/2)n^{2/(d+2)}$ for all $k \in K_n$, which implies that

$$\#\{t \leq n: I_t \in B_k^{(j_1, \dots, j_d)}\} \geq (C_1/2) 2^{j_1 + \dots + j_d} n^{2/(d+2)}.$$

Furthermore, it follows from Lemma 5.2 that for some $C < \infty$

$$E \left[\sup \left\{ \left| \sum_{t=1}^n \varepsilon_t \mathbb{1}(I_t \in (y, x_{k+1}]) \right| : y \in B_k^{(j_1+1, \dots, j_d+1)} \right\} \right] \leq C 2^{(j_1 + \dots + j_d)/2} n^{1/(d+2)}. \quad (5.7)$$

Therefore, we obtain that

$$\begin{aligned} & \sup_{x \in B_k} E \left[\sup_{y \leq x_k} |\text{Av}_\varepsilon((y, x_{k+1}])| \mathbb{1}_{A_n} \right] \\ & \leq \sum_{j_1, \dots, j_d \geq 0} E \left[\frac{\sup \left\{ \left| \sum_{t=1}^n \varepsilon_t \mathbb{1}(I_t \in (y, x_{k+1}]) \right| : y \in B_k^{(j_1+1, \dots, j_d+1)} \right\}}{\#\{t \leq n: I_t \in B_k^{(j_1, \dots, j_d)}\}} \mathbb{1}_{A_n} \right] \\ & = O \left(n^{-1/(d+2)} \sum_{j_1, \dots, j_d \geq 0} 2^{-(j_1 + \dots + j_d)/2} \right) = O(n^{-1/(d+2)}). \end{aligned}$$

This yields, in conjunction with (5.5) and (5.6), that (5.4a) holds. The proof of (5.4b) is completely analogous and therefore it is omitted.

Proof of (ii): We have that

$$\begin{aligned}
& E \left[\int_{[0,1]^d} |\widehat{f}_n(x) - f(x)| \lambda^d(dx) \right] \\
&= E \left[\int_{D_n} |\widehat{f}_n(x) - f(x)| \lambda^d(dx) \mathbb{1}_{A_n} \right] \\
&\quad + E \left[\int_{D_n} |\widehat{f}_n(x) - f(x)| \lambda^d(dx) \mathbb{1}_{A_n^c} \right] \\
&\quad + E \left[\int_{[0,1]^d \setminus D_n} |\widehat{f}_n(x) - f(x)| \lambda^d(dx) \right].
\end{aligned} \tag{5.8}$$

By (5.4a) and (5.4b), the first term on the right-hand side is of order $O(n^{-1/(d+2)})$. Since $\underline{f} \leq \widehat{f}_n(x), f(x) \leq \bar{f}$ for all $x \in [0,1]^d$ the integrands of the integrals on the right-hand side are bounded by $\bar{f} - \underline{f}$. Since $P(A_n^c) = O(n^{-1/(d+2)})$ by (5.2) and $\lambda^d([0,1]^d \setminus D_n) = O(n^{-1/(d+2)})$ we conclude that the last two terms on the right-hand side of (5.8) are also of order $O(n^{-1/(d+2)})$, as required. \square

Proof of Theorem 3.1. The proof of this theorem is largely the same as that of Theorem 2.1. First we split the set $\widetilde{D} = \{0, \dots, K\}^{d_1} \times [0,1]^{d_2}$ into subsets which are adapted to our assumption (A3)(i) on the distribution of the covariates. Recall that $\widetilde{h}_n = 1/\widetilde{M}_n$, where $\widetilde{M}_n = \lceil n^{1/(d_2+2)} \rceil$. Let, for multi-indexes $k = (k_1, \dots, k_d) \in \widetilde{K}_n = \{0, \dots, K\}^{d_1} \times \{1, \dots, \widetilde{M}_n\}^{d_2}$,

$$\widetilde{B}_k = \{(k_1, \dots, k_{d_1})'\} \times (G_1^{-1}((k_{d_1+1}-1)\widetilde{h}_n), G_1^{-1}(k_{d_1+1}\widetilde{h}_n)) \times \dots \times (G_{d_2}^{-1}((k_d-1)\widetilde{h}_n), G_{d_2}^{-1}(k_d\widetilde{h}_n)).$$

As in the proof of Theorem 2.1, we define a set which describes a ‘‘regular’’ behavior of the explanatory variables by

$$\widetilde{A}_n = \{\omega: \#\{t \leq n: I_t(\omega) \in \widetilde{B}_k\} \geq C_4 n^{2/(d_2+2)} \quad \forall k \in \widetilde{K}_n\}, \tag{5.9}$$

where C_4 is some positive constant.

Proof of (i): In what follows we show that

$$E \left[\int_{\widetilde{D}_n} (\widehat{f}_n(x) - f(x))_+ \nu(dx) \mathbb{1}_{\widetilde{A}_n} \right] = O(n^{-1/(d_2+2)}), \tag{5.10a}$$

and

$$E \left[\int_{\widetilde{D}_n} (\widehat{f}_n(x) - f(x))_- \nu(dx) \mathbb{1}_{\widetilde{A}_n} \right] = O(n^{-1/(d_2+2)}). \tag{5.10b}$$

Since by Lemma 5.3

$$P(\tilde{A}_n) \xrightarrow{n \rightarrow \infty} 1$$

we then obtain that assertion (i) holds true. We define grid points

$$\begin{aligned} \bar{x}_k &= \\ & (k_1, \dots, k_{d_1}, G_1^{-1}((k_{d_1+1} + 1)h_n), \dots, G_{d_2}^{-1}((k_d + 1)h_n))', \\ \underline{x}_k &= \\ & (k_1, \dots, k_{d_1}, G_1^{-1}((k_{d_1+1} - 1)h_n), \dots, G_{d_2}^{-1}((k_d - 1)h_n))'. \end{aligned}$$

We have, for all $x \in \tilde{B}_k$,

$$\begin{aligned} (\widehat{f}_n(x) - f(x))_+ \mathbb{1}_{A_n} &\leq \left(\sup_{y \leq x_k} \text{Av}_Y((y, \bar{x}_k]) - f(x) \right)_+ \mathbb{1}_{A_n} \\ &\leq \sup_{y \leq x_k} |\text{Av}_\varepsilon((y, \bar{x}_k])| + (f(\bar{x}_k) - f(x)). \end{aligned} \quad (5.11)$$

We apply Lemma 5.1 to $\tilde{f}(\tilde{x}_1, \dots, \tilde{x}_{d_2}) = f(k_1, \dots, k_{d_1}, G_1^{-1}(\tilde{x}_1), \dots, G_{d_2}^{-1}(\tilde{x}_{d_2}))$, $(\tilde{x}_1, \dots, \tilde{x}_{d_2})' \in [0, 1]^{d_2}$, with $M = \tilde{M}_n$. Since $\nu^d(\tilde{B}_k) = \tilde{h}_n^{d_2} = O(n^{-d_2/(d_2+2)})$ we obtain for the bias that

$$\begin{aligned} &\sum_{k: 1 \leq k_{d_1+1}, \dots, k_d < \tilde{M}_n} \int_{\tilde{B}_k} (f(\bar{x}_k) - f(x)) \nu(dx) \\ &\leq \sum_{k: 1 \leq k_{d_1+1}, \dots, k_d < \tilde{M}_n} (f(\bar{x}_k) - f(\underline{x}_k)) \nu(\tilde{B}_k) \\ &= O(\tilde{M}_n^{d_2-1} \tilde{h}_n^{d_2}) = O(n^{-1/(d_2+2)}). \end{aligned} \quad (5.12)$$

We define again a dyadic scheme of nested hyperrectangles: For $j_1, \dots, j_{d_2} \geq 0$,

$$\tilde{B}_k^{(j_1, \dots, j_{d_2})} = \{(k_1, \dots, k_{d_1})'\} \times ((k_{d_1+1} + 1 - 2^{j_1})\tilde{h}_n, (k_{d_1+1} + 1)\tilde{h}_n] \times \dots \times ((k_d + 1 - 2^{j_{d_2}})\tilde{h}_n, (k_d + 1)\tilde{h}_n]$$

and

$$\begin{aligned} \tilde{B}_{k,0}^{(j_1, \dots, j_{d_2})} &= \{0, \dots, k_1\} \times \dots \times \{0, \dots, k_{d_1}\} \times ((k_{d_1+1} + 1 - 2^{j_1})\tilde{h}_n, (k_{d_1+1} + 1)\tilde{h}_n] \times \\ &\quad \dots \times ((k_d + 1 - 2^{j_{d_2}})\tilde{h}_n, (k_d + 1)\tilde{h}_n]. \end{aligned}$$

Since $\{y \leq x_k\} \subseteq \bigcup_{j_1, \dots, j_{d_2} \geq 0} \tilde{B}_{k,0}^{(j_1+1, \dots, j_{d_2}+1)} \setminus \tilde{B}_{k,0}^{(j_1, \dots, j_{d_2})}$ and since $y \in \tilde{B}_{k,0}^{(j_1+1, \dots, j_{d_2}+1)} \setminus \tilde{B}_{k,0}^{(j_1, \dots, j_{d_2})}$ implies that $(y, \bar{x}_k] \supseteq \tilde{B}_k^{(j_1, \dots, j_{d_2})}$ we obtain, for all $x \in \tilde{B}_k$,

$$\sup_{y \leq x_k} |\text{Av}_\varepsilon((y, \bar{x}_k])| \mathbb{1}_{A_n} \leq \sum_{j_1, \dots, j_{d_2} \geq 0} \frac{\sup \left\{ \left| \sum_{t=1}^n \varepsilon_t \mathbb{1}(I_t \in (y, \bar{x}_k]) \right| : y \in \tilde{B}_{k,0}^{(j_1+1, \dots, j_{d_2}+1)} \setminus \tilde{B}_{k,0}^{(j_1, \dots, j_{d_2})} \right\}}{\#\{t \leq n: I_t \in \tilde{B}_k^{(j_1, \dots, j_{d_2})}\}} \mathbb{1}_{A_n}.$$

Recall that if the event \tilde{A}_n occurs, then $\#\{t \leq n: I_t \in \tilde{B}_k\} \geq C_4 n^{2/(d_2+2)}$ for all $k \in \tilde{K}_n$, which implies that

$$\#\{t \leq n: I_t \in \tilde{B}_k^{(j_1, \dots, j_{d_2})}\} \geq C_4 2^{j_1 + \dots + j_{d_2}} n^{2/(d_2+2)}.$$

Furthermore, it follows from Lemma 5.4 that for some $C < \infty$

$$E \left[\sup \left\{ \left| \sum_{t=1}^n \varepsilon_t \mathbb{1}(I_t \in (y, \bar{x}_k]) \right| : y \in \tilde{B}_{k,0}^{(j_1+1, \dots, j_{d_2}+1)} \right\} \right] \leq C 2^{(j_1 + \dots + j_{d_2})/2} n^{1/(d_2+2)}. \quad (5.13)$$

Therefore, we obtain that

$$\begin{aligned} & \sup_{x \in \tilde{B}_k} E \left[\sup_{y \leq x_k} |\text{Av}_\varepsilon((y, \bar{x}_k])| \mathbb{1}_{\tilde{A}_n} \right] \\ & \leq \sum_{j_1, \dots, j_{d_2} \geq 0} E \left[\frac{\sup \left\{ \left| \sum_{t=1}^n \varepsilon_t \mathbb{1}(I_t \in (y, \bar{x}_k]) \right| : y \in \tilde{B}_{k,0}^{(j_1+1, \dots, j_{d_2}+1)} \right\}}{\#\{t \leq n: I_t \in \tilde{B}_k^{(j_1, \dots, j_{d_2})}\}} \mathbb{1}_{\tilde{A}_n} \right] \\ & = O \left(n^{-1/(d_2+2)} \sum_{j_1, \dots, j_{d_2} \geq 0} 2^{-(j_1 + \dots + j_{d_2})/2} \right) = O(n^{-1/(d_2+2)}). \end{aligned}$$

This yields, in conjunction with (5.11) and (5.12), that (5.10a) holds. The proof of (5.10b) is completely analogous and therefore it is omitted.

Proof of (ii): The second assertion follows in the same way as that of Theorem 2.1. \square

5.2. Some auxiliary results.

Lemma 5.1. *Suppose that $f: [0, 1]^d \rightarrow \mathbb{R}$ is isotonic and let, for $M \in \mathbb{N}$ and $k = (k_1, \dots, k_d)$, $x_k = (k_1/M, \dots, k_d/M)$. Then*

$$\sum_{k: 0 < k_1, \dots, k_d < M} (f(x_{k+1}) - f(x_{k-1})) \leq 2d M^{d-1} (f(1, \dots, 1) - f(0, \dots, 0)).$$

Proof of Lemma 5.1. Let $\mathcal{I}_0 = \{k: 0 < k_1, \dots, k_d < M \text{ and } k_j = 1 \text{ for at least one } j\}$.

We estimate the sum by considering the main and minor diagonals as follows:

$$\begin{aligned} \sum_{k: 0 < k_1, \dots, k_d < M} (f(x_{k+1}) - f(x_{k-1})) &= \sum_{k \in \mathcal{I}_0} \sum_{i \geq 0} (f(x_{k+(i+1)\mathbf{1}}) - f(x_{k+(i-1)\mathbf{1}})) \\ &\leq \#\mathcal{I}_0 \cdot 2 \left(\sup_x \{f(x)\} - \inf_x \{f(x)\} \right). \end{aligned}$$

The assertion of the lemma follows because $\#\mathcal{I}_0 \leq dM^{d-1}$. \square

Lemma 5.2. *Suppose that the assumptions of Theorem 2.1 hold true. Then, for arbitrary $\underline{z} \leq \bar{z}$ with $[\underline{z}, \bar{z}] \subseteq [0, 1]^d$ and some $\bar{C} < \infty$,*

$$E \left[\sup_{z: \underline{z} \leq z \leq \bar{z}} \left| \frac{1}{\sqrt{n}} \sum_{t=1}^n \varepsilon_t \mathbb{1}(I_t \in (z, \bar{z}]) \right| \right] \leq \bar{C} \sqrt{P(I_1 \in (\underline{z}, \bar{z}])} \quad (5.14a)$$

and

$$E \left[\sup_{z: \underline{z} \leq z \leq \bar{z}} \left| \frac{1}{\sqrt{n}} \sum_{t=1}^n \varepsilon_t \mathbb{1}(I_t \in [\underline{z}, z]) \right| \right] \leq \bar{C} \sqrt{P(I_1 \in [\underline{z}, \bar{z}])}. \quad (5.14b)$$

Proof of Lemma 5.2. We prove only (5.14a) since the proof of (5.14b) is completely analogous. One of the main tools which is used is given by Bickel and Wichura (1971, Thm. 1). For this purpose, we adopt some notation from there. A block B in $[\underline{z}, \bar{z}]$ is a subset of $[\underline{z}, \bar{z}]$ of the form $(u, v) = (u_1, v_1] \times \dots \times (u_d, v_d]$. For $p \in \{1, \dots, d\}$, the p th face of $B = (u, v)$ is $(u_1, v_1] \times \dots \times (u_{p-1}, v_{p-1}] \times (u_{p+1}, v_{p+1}] \times \dots \times (u_d, v_d]$. Disjoint blocks B and C are p -neighbors if they are abut and have the same p th face; they are neighbors if they are p -neighbors for some p . (For example, $(u_1, v_1] \times \dots \times (u_{p-1}, v_{p-1}] \times (\tilde{u}, \tilde{v}] \times (u_{p+1}, v_{p+1}] \times \dots \times (u_d, v_d]$ and $(u_1, v_1] \times \dots \times (u_{p-1}, v_{p-1}] \times (\tilde{v}, \tilde{w}] \times (u_{p+1}, v_{p+1}] \times \dots \times (u_d, v_d]$ are p -neighbors if $0 \leq \tilde{u} < \tilde{v} < \tilde{w} \leq 1$.) For each block $B = (u, v]$, let

$$X(B) = \frac{1}{\sqrt{n}} \sum_{t=1}^n \varepsilon_t \mathbb{1}(I_t \in B).$$

In what follows we show that condition (2) in Bickel and Wichura (1971, Thm. 1) is fulfilled. To this end, let B and C be arbitrary neighboring blocks in $[\underline{z}, \bar{z}]$. We will estimate the expected value of the term

$$|X(B)|^2|X(C)|^2 = \frac{1}{n^2} \sum_{t_1, t_2, t_3, t_4=1}^n \mathbb{1}(I_{t_1} \in B) \mathbb{1}(I_{t_2} \in B) \mathbb{1}(I_{t_3} \in C) \mathbb{1}(I_{t_4} \in C) \varepsilon_{t_1} \varepsilon_{t_2} \varepsilon_{t_3} \varepsilon_{t_4}.$$

Since B and C are disjoint sets it follows that

$$\mathbb{1}(I_{t_1} \in B) \mathbb{1}(I_{t_2} \in B) \mathbb{1}(I_{t_3} \in C) \mathbb{1}(I_{t_4} \in C) = 0,$$

if $\{t_1, t_2\} \cap \{t_3, t_4\} \neq \emptyset$. Therefore, and by independence of $(I'_1, \varepsilon_1)', \dots, (I'_n, \varepsilon_n)'$,

$$\begin{aligned} & E[|X(B)|^2|X(C)|^2] \\ &= \frac{1}{n^2} \sum_{(t_1, \dots, t_4): \{t_1, t_2\} \cap \{t_3, t_4\} = \emptyset} E[\mathbb{1}(I_{t_1} \in B) \mathbb{1}(I_{t_2} \in B) \varepsilon_{t_1} \varepsilon_{t_2}] E[\mathbb{1}(I_{t_3} \in C) \mathbb{1}(I_{t_4} \in C) \varepsilon_{t_3} \varepsilon_{t_4}]. \end{aligned}$$

Furthermore, again by independence of $(I'_1, \varepsilon_1)', \dots, (I'_n, \varepsilon_n)'$, and since

$$E[\mathbb{1}(I_s \in B) \mathbb{1}(I_t \in B) \varepsilon_s \varepsilon_t] = E[\mathbb{1}(I_s \in C) \mathbb{1}(I_t \in C) \varepsilon_s \varepsilon_t] = 0,$$

if $s \neq t$, we obtain that

$$\begin{aligned} E[|X(B)|^2|X(C)|^2] &= \frac{1}{n^2} \sum_{s \neq t} E[\mathbb{1}(I_s \in B) \varepsilon_s^2] E[\mathbb{1}(I_t \in C) \varepsilon_t^2] \\ &\leq \bar{\sigma}_\varepsilon^4 P(I_1 \in B) P(I_1 \in C). \end{aligned} \tag{5.15}$$

Let $m(B, C) = \min\{|X(B)|, |X(C)|\}$. From (5.15) we obtain by Markov's inequality that

$$P(m(B, C) \geq \nu) \leq \frac{E[m(B, C)^4]}{\nu^4} \leq \frac{E[|X(B)|^2|X(C)|^2]}{\nu^4} \leq \nu^{-4} \mu(B) \mu(C)$$

for all $\nu > 0$ and the measure $\mu(\cdot) = \bar{\sigma}_\varepsilon^2 P^{I_1}(\cdot)$. Hence, condition (2) of Bickel and Wichura (1971, Thm. 1) is fulfilled and it follows that

$$P\left(\sup_{z: \underline{z} \leq z \leq \bar{z}} \left| \frac{1}{\sqrt{n}} \sum_{t=1}^n \varepsilon_t \mathbb{1}(I_t \in (\underline{z}, z]) \right| \geq \nu\right) \leq \tilde{C} \nu^{-4} \mu((\underline{z}, \bar{z}])^2,$$

for all $\nu > 0$ and some $\tilde{C} < \infty$. This, however, implies that

$$\begin{aligned}
& E \left[\sup_{z: \underline{z} \leq z \leq \bar{z}} \left| \frac{1}{\sqrt{n}} \sum_{t=1}^n \varepsilon_t \mathbb{1}(I_t \in (z, \bar{z}]) \right| \right] \\
&= \int_0^\infty P \left(\sup_{z: \underline{z} \leq z \leq \bar{z}} \left| \frac{1}{\sqrt{n}} \sum_{t=1}^n \varepsilon_t \mathbb{1}(I_t \in (z, \bar{z}]) \right| \geq \lambda \right) d\lambda \\
&\leq \sqrt{\mu((\underline{z}, \bar{z}])} + \int_{\sqrt{\mu((\underline{z}, \bar{z}])}}^\infty P \left(\sup_{z: \underline{z} \leq z \leq \bar{z}} \left| \frac{1}{\sqrt{n}} \sum_{t=1}^n \varepsilon_t \mathbb{1}(I_t \in (z, \bar{z}]) \right| \geq \lambda \right) d\lambda \\
&\leq \sqrt{\mu((\underline{z}, \bar{z}])} + 3\tilde{C} \sqrt{\mu((\underline{z}, \bar{z}])},
\end{aligned}$$

which proves the assertion of the lemma. \square

Lemma 5.3. *Suppose that Assumptions (A2) and (A3) hold true. Let \tilde{A}_n be defined as in the proof of Theorem 3.1. Then, for sufficiently small $C_4 > 0$ in (5.9),*

$$P(\tilde{A}_n^c) = O(n^{-1/(d_2+2)}).$$

Proof of Lemma 5.3. We will show that

$$\max_{k \in \tilde{K}_n} \left\{ P \left(\sum_{t=1}^n P(I_t \in \tilde{B}_k) - \mathbb{1}(I_t \in \tilde{B}_k) \geq \frac{C_1 n^{2/(d_2+2)}}{2} \right) \right\} = O(n^{-(d_2+1)/(d_2+2)}). \quad (5.16)$$

Let $\mu_n = C_1 n^{2/(d_2+2)}/8$ and, for arbitrary $k \in \tilde{K}_n$, $\eta_t = \mathbb{1}(I_t \in \tilde{B}_k) - P(I_t \in \tilde{B}_k)$. It follows from the Fuk-Nagaev-type inequality (I.6) of Rio (2000, page 4) that, for all $\kappa \geq 1$,

$$P \left(\left| \sum_{t=1}^n \eta_t \right| \geq 4\mu_n \right) \leq \left(1 + \frac{\mu_n^2}{\kappa s_n^2} \right)^{-\kappa/2} + \frac{n \alpha([\mu_n/\kappa])}{\mu_n}, \quad (5.17)$$

where

$$s_n^2 = \sum_{s,t=1}^n |\text{cov}(\eta_s, \eta_t)|.$$

Since $\alpha([\mu_n/\kappa]) = O(n^{-(2d_2+1)/(d_2+2)})$ we obtain that

$$\frac{n \alpha([\mu_n/\kappa])}{\mu_n} = O(n^{-(d_2+1)/(d_2+2)}),$$

that is, the second term on the right-hand side of (5.17) is of the required order.

It remains to estimate s_n^2/μ_n^2 . We obtain from a covariance inequality for strong mixing processes (see e.g. Bradley (2007, Corollary 10.16)) that

$$|\text{cov}(\eta_t, \eta_{t+r})| \leq 4 \alpha(r) \|\eta_t\|_\infty \|\eta_{t+r}\|_\infty \leq 4 \alpha(r).$$

If $d_2 = 0$, then

$$s_n^2 \leq 2n \sum_{r=1}^{n-1} 4 \alpha(r) = O(n^{3/2}),$$

which implies that

$$\frac{s_n^2}{\mu_n^2} = O(n^{-1/2}). \quad (5.18)$$

If $d_2 \geq 1$, then we distinguish between the two cases of covariates without and with a trend component. In the first case, we obtain from the upper bound in (A3)(i) and (ii) that, for all t, r with $1 \leq t \leq t+r \leq n$,

$$|\text{cov}(\eta_t, \eta_{t+r})| = \begin{cases} O(n^{-d_2/(d_2+2)}) & \text{if } 0 \leq r < d, \\ O(n^{-2d_2/(d_2+2)}) & \text{if } r \geq d \end{cases}.$$

Therefore, with $N_n = \lceil n^{d_2/(d_2+2)} \rceil$,

$$\begin{aligned} s_n^2 &\leq \sum_{s,t: |s-t| < d} |\text{cov}(\eta_s, \eta_t)| + \sum_{s,t: d \leq |s-t| \leq N_n} |\text{cov}(\eta_s, \eta_t)| + 2n \sum_{r=N_n+1}^{n-1} 4 \alpha(r) \\ &= O(n^{2/(d_2+2)}) + O(n^{2/(d_2+2)}) + O(n N_n^{1/2-d_2}) \\ &= O(n^{2/(d_2+2)} + n^{[2+(3/2-d_2)d_2]/(d_2+2)}), \end{aligned}$$

which implies that

$$\frac{s_n^2}{\mu_n^2} = O(n^{-2/(d_2+2)} + n^{[(3/2-d_2)d_2-2]/(d_2+2)}). \quad (5.19)$$

In the case with trend, we get from (A3)(i) that

$$|\text{cov}(\eta_t, \eta_{t+r})| = \begin{cases} O(n^{-(d_2-1)/(d_2+2)}) & \text{if } 0 \leq r < d-1, \\ O(n^{-2(d_2-1)/(d_2+2)}) & \text{if } r \geq d-1 \end{cases}.$$

On the other hand, we see that $I_t = (\tilde{I}'_t, t/n)' \notin \tilde{B}_k$, and therefore $\eta_t = 0$ if $t \notin I_{n,k} := ((k_d - 1)nh_n, k_dnh_n]$. Hence, here with $N_n = \lceil n^{(d_2-1)/(d_2+2)} \rceil$,

$$\begin{aligned} s_n^2 &\leq \sum_{s,t \in I_{n,k}} |\text{cov}(\eta_s, \eta_t)| \\ &\leq \sum_{(s,t) \in I_{n,k}: |s-t| < d-1} |\text{cov}(\eta_s, \eta_t)| + \sum_{(s,t) \in I_{n,k}: d-1 \leq |s-t| \leq N_n} |\text{cov}(\eta_s, \eta_t)| + nh_n \sum_{r=N_n+1}^{n-1} 4\alpha(r) \\ &= O(n^{2/(d_2+2)}) + O(n^{2/(d_2+2)}) + O(n^{2/(d_2+2)} n^{(3/2-d_2)(d_2-1)/(d_2+2)}), \end{aligned}$$

which implies that

$$\frac{s_n^2}{\mu_n^2} = O(n^{-2/(d_2+2)}). \quad (5.20)$$

We see from (5.18), (5.19) and (5.20) that in all cases the term $(1 + \mu_n^2/(\kappa s_n^2))^{-1}$ is of order $O(n^{-\gamma})$, for some $\gamma > 0$. Choosing $\kappa > 2d_2/\gamma$ we see that (5.16) follows from (5.17), which completes the proof. \square

Lemma 5.4. *Suppose that the assumptions of Theorem 3.1 hold true. Define $\rho_n = n^{-1} \sum_{t=1}^n P^{I_{n,t}}$. Then, for arbitrary $\underline{z} \leq \bar{z}$ with $[\underline{z}, \bar{z}] \subseteq \mathbb{N}_0^{d_1} \times \mathbb{R}^{d_2}$ and some $\bar{C} < \infty$,*

$$E \left[\sup_{\underline{z}: \underline{z} \leq z \leq \bar{z}} \left| \frac{1}{\sqrt{n}} \sum_{t=1}^n \varepsilon_t \mathbb{1}(I_t \in (z, \bar{z})) \right| \right] \leq \bar{C} \sqrt{\rho_n([\underline{z}, \bar{z}])} \quad (5.21a)$$

and

$$E \left[\sup_{\underline{z}: \underline{z} \leq z \leq \bar{z}} \left| \frac{1}{\sqrt{n}} \sum_{t=1}^n \varepsilon_t \mathbb{1}(I_t \in [\underline{z}, z]) \right| \right] \leq \bar{C} \sqrt{\rho_n([\underline{z}, \bar{z}])}. \quad (5.21b)$$

Proof of Lemma 5.4. The proof is pretty much the same as that of Lemma 5.2. Since we impose condition (A3)(ii), we have only a bound for the conditional probability $P(I_t \in C \mid I_1, \dots, I_{t-d}, \varepsilon_1, \dots, \varepsilon_{t-d})$ but not for $P(I_t \in C \mid I_1, \dots, I_{t-1}, \varepsilon_1, \dots, \varepsilon_{t-1})$ at our disposal. In view of this, we consider first the d -thinned partial sums

$$X_i(B) = \frac{1}{\sqrt{n}} \sum_{s: 1 \leq sd+i \leq n} \varepsilon_{sd+i} \mathbb{1}(I_{sd+i} \in B),$$

for $i = 1, \dots, d$, instead of the full partial sums. In analogy to (5.15) in the proof of Lemma 5.2 we show that, for any neighboring blocks B and C in $[\underline{z}, \bar{z}]$ and any $i \in \{1, \dots, d\}$,

$$E[|X_i(B)|^2 |X_i(C)|^2] \leq \tilde{C} \rho_n(B) \rho_n(C), \quad (5.22)$$

for some $\tilde{C} < \infty$.

As in the independent regressors case, we consider again, for arbitrary neighbored blocks B and C and arbitrary $t_1, t_2, t_3, t_4 \in \{1, \dots, n\}$, the terms $E[\mathbb{1}(I_{t_1} \in B)\mathbb{1}(I_{t_2} \in B)\mathbb{1}(I_{t_3} \in C)\mathbb{1}(I_{t_4} \in C)\varepsilon_{t_1}\varepsilon_{t_2}\varepsilon_{t_3}\varepsilon_{t_4}]$. Since B and C are disjoint sets it follows as before that $\mathbb{1}(I_{t_1} \in B)\mathbb{1}(I_{t_2} \in B)\mathbb{1}(I_{t_3} \in C)\mathbb{1}(I_{t_4} \in C) = 0$ provided that $\{t_1, t_2\} \cap \{t_3, t_4\} \neq \emptyset$. This implies that the above expectation is equal to 0. Moreover, if the largest index appears only once, then the expectation also vanishes since, by (A2)(i), $E(\varepsilon_t | I_1, \dots, I_t, \varepsilon_1, \dots, \varepsilon_{t-1}) = 0$. Therefore, we have to examine in more detail two cases: $1 \leq t_1, t_2 < t_3 = t_4 \leq n$ and $1 \leq t_3, t_4 < t_1 = t_2 \leq n$. Hence we obtain that

$$\begin{aligned} & E[|X_i(B)|^2 |X_i(C)|^2] \\ &= \frac{1}{n} \sum_{t: d < td+i \leq n} E \left[\left(\frac{1}{\sqrt{n}} \sum_{s=0}^{t-1} \mathbb{1}(I_{sd+i} \in B) \varepsilon_{sd+i} \right)^2 \mathbb{1}(I_{td+i} \in C) \varepsilon_{td+i}^2 \right] \\ &\quad + \frac{1}{n} \sum_{t: d < td+i \leq n} E \left[\left(\frac{1}{\sqrt{n}} \sum_{s=0}^{t-1} \mathbb{1}(I_{sd+i} \in C) \varepsilon_{sd+i} \right)^2 \mathbb{1}(I_{td+i} \in B) \varepsilon_{td+i}^2 \right] \\ &\leq \bar{\sigma}_\varepsilon^2 \frac{1}{n} \sum_{t: d < td+i \leq n} E \left[\left(\frac{1}{\sqrt{n}} \sum_{s=0}^{t-1} \mathbb{1}(I_{sd+i} \in B) \varepsilon_{sd+i} \right)^2 P(I_{td+i} \in C | I_1, \dots, I_{(t-1)d+i}, \varepsilon_1, \dots, \varepsilon_{(t-1)d+i}) \right] \\ &\quad + \bar{\sigma}_\varepsilon^2 \frac{1}{n} \sum_{t: d < td+i \leq n} E \left[\left(\frac{1}{\sqrt{n}} \sum_{s=0}^{t-1} \mathbb{1}(I_{sd+i} \in C) \varepsilon_{sd+i} \right)^2 P(I_{td+i} \in B | I_1, \dots, I_{(t-1)d+i}, \varepsilon_1, \dots, \varepsilon_{(t-1)d+i}) \right] \\ &\leq \tilde{C} \rho_n(B) \rho_n(C), \end{aligned}$$

as required. Using (5.22) to estimate (5.15), similar to the proof of Lemma 5.2, we obtain in the same manner that

$$E \left[\sup_{z: \underline{z} \leq z \leq \bar{z}} |X_i((z, \bar{z}))| \right] \leq \bar{C} \sqrt{\rho_n((\underline{z}, \bar{z}))}.$$

Finally, summing up over $i = 1, \dots, d$ we obtain (5.21a). The proof of (5.21b) is analogous and therefore it is omitted. \square

Acknowledgment . We thank the Associate Editor R. Balan and two anonymous reviewers for many constructive comments that improved earlier versions of this work. This research was partly funded by the German Research Foundation DFG, project NE 606/2-2 and by the Volkswagen Foundation (Professorinnen für Niedersachsen des Niedersächsischen Vorab).

REFERENCES

- ANEVSKI, D. AND HÖSSJER, O. (2006). A general asymptotic scheme for inference under order restrictions. *Ann. Statist.* **34**, 1874–1930.
- BICKEL, P. J. AND WICHURA, M. J. (1971). Convergence criteria for multiparameter stochastic processes and some applications. *Ann. Math. Statist.* **42**, 1656–1670.
- BRADLEY, R. C. (2007). *Introduction to Strong Mixing Conditions, Volume I*. Kendrick Press.
- BRUNK, H. D. (1955). Maximum likelihood estimates of monotone parameters. *Ann. Math. Statist.* **26**, 607–616.
- CHATTERJEE, S., GUNTUBOYINA, A., AND SEN, B. (2015). On risk bounds in isotonic regression and other shape restricted regression problems. *Ann. Statist.* **43**, 1774–1800.
- CHEN, Y. AND SAMWORTH, R. J. (2016). Generalized additive and index models with shape constraints. *J. R. Statist. Soc. B.* **78**, 729–754.
- CHERNOZHUKOV, V. AND FERNÁNDEZ-VAL, I., AND GALICHON, A. (2009). Improving point and interval estimators of monotone functions by rearrangement. *Biometrika* **96**, 559–575.
- CHRISTOPEIT, N. AND TOSSTORFF, G. (1987). Strong consistency of least-squares estimators in the monotone regression model with stochastic regressors. *Ann. Statist.* **15**, 568–586.

- DAHLHAUS, R. (1997). Fitting time series models to nonstationary processes. *Ann. Statist.* **25**, 1–37.
- DAHLHAUS, R. AND NEUMANN, M. H. (2001). Locally adaptive fitting of semiparametric models to nonstationary time series. *Stochastic Processes and Applications* **91**, 277–308.
- DAOUIA, A. AND PARK, B. U. (2013). On projection-type estimators of multivariate isotonic unctons. *Scand. J. Stat.*, **40**, 363–386.
- DEDECKER, J., MERLEVÈDE, F., AND PELIGRAD, M. (2011). Invariance principles for linear processes with application to isotonic regression. *Bernoulli* **17**, 88–113.
- DENG, H. AND ZHANG, C.-H. (2020). Isotonic regression in multi-dimensional spaces and graphs. *arXiv:1812.08944v2*
- DETTE, H., NEUMEYER, N., AND PILZ, K. F. (2006). A simple nonparametric estimator of a strictly monotone regression function. *Bernoulli* **12**, 469–490.
- DOUKHAN, P. (1994). *Mixing: Properties and Examples*. Lecture Notes in Statistics **85**. Springer-Verlag, Berlin, Heidelberg.
- DUROT, C. (2002). Sharp asymptotics for isotonic regression. *Probab. Theory Relat. Fields* **122**, 222–240.
- ESCANCIANO, J. C. (2006). Goodness-of-fit tests for linear and nonlinear time series models. *J. Amer. Statist. Assoc.* **101**, 531–541.
- FAN, J. AND GIJBELS, I. (1996). *Local Polynomial Modelling and its Applications*. London: Chapman & Hall.
- FOKIANOS, K., RAHBEK, A., AND TJØSTHEIM, D. (2009). Poisson autoregression, *J. Amer. Statist. Assoc.* **104**, 1430–1439.
- FRANCO, C. AND ZAKOÏAN, J.-M. (2010). *GARCH models: Structure, Statistical Inference and Financial Applications*. UK: Wiley.
- GAO, F. AND WELLNER, J. A. (2007). Entropy estimate for high-dimensional monotonic functions. *J. Mult. Anal.* **98**, 1751–1764.

- HAN, Q., WANG, T., CHATTERJEE, S., AND SAMWORTH R. J. (2019). Isotonic regression in general dimensions. *Ann. Statist.* **47**, 2440–2471.
- HAN, Q. AND ZHANG, C.-H. (2019). Limit distribution theory for multiple isotonic regression. *arXiv:1905.12825v2*
- HANSON, D. L., PLEDGER, G., AND WRIGHT, F. T. (1973). On consistency in monotonic regression. *Ann. Statist.* **1**, 401–421.
- HÄRDLE, W. (1990). *Applied Nonparametric Regression*. Cambridge; New York; New Rochelle: Cambridge University Press.
- INTERNATIONAL RECOVERY PLAN FOR THE WHOOPING CRANE (*Grus Americana*), THIRD REVISION (2007). *Endangered Species Bulletins and Technical Reports (US Fish and Wildlife Service)*. Paper 45.
- MUKARJEE, H. AND STERN, S. (1994). Feasible nonparametric estimation of multiargument monotone functions. *J. Amer. Stat. Assoc.*, **89**, 77–80.
- RIO, E. (2000). *Théorie asymptotique des processus aléatoires faiblement dépendants*. Mathématiques & Applications 31, Springer.
- ROBERTSON, T. AND WRIGHT, F. T. (1975). Consistency in generalized isotonic regression. *Ann. Statist.* **3**, 350–362.
- ROBERTSON, T., WRIGHT, F. T., AND DYKSTRA, R. L. (1988). *Order Restricted Statistical Inference*. Wiley, New York.
- SAMPSON, A. R., SINGH, H., AND WHITAKER, L. R. (2003). Order restricted estimators: some bias results. *Statist. Probab. Lett.* **61**, 299–308.
- SHUMWAY, R. H. AND STOFFER, D. S. (2011) *Time Series Analysis and its Applications*. Third ed. Springer, New York.
- STONE, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10**, 1040–1053.
- WU, J., MEYER, M. C., AND OPSOMER, J. D. (2015). Penalized isotonic regression. *J. Statist. Plann. Inference* **161**, 12–24.
- ZHANG, C.-H. (2002). Risk bounds in isotonic regression. *Ann. Statist.* **30**, 528–555.