Characterization of the least squares estimator: Mis-specified multivariate isotonic regression model with dependent errors

Article *in* Theory of Probability and Mathematical Statistics · April 2023

DOI: 10.1090/tpms/1210

Pramita Bagchi
Ruhr University Bochum
41 PUBLICATIONS **354** CITATIONS

Subhra Sankar Dhar
Indian Institute of Technology Kanpur
59 PUBLICATIONS **325** CITATIONS

Theor. Probability and Math. Statist.
No. 0, XXXX, Pages 000–000

CHARACTERIZATION OF THE LEAST SQUARES ESTIMATOR : MIS-SPECIFIED MULTIVARIATE ISOTONIC REGRESSION MODEL WITH DEPENDENT ERRORS

PRAMITA BAGCHI AND SUBHRA SANKAR DHAR

ABSTRACT. This article investigates some nice properties of the least squares estimator of multivariate isotonic regression function (denoted as LSEMIR), when the model is mis-specified, and the errors are β-mixing stationary random variables. Under mild conditions, it is observed that the least squares estimator converges uniformly to a certain monotone function, which is closest to the original function in an appropriate sense.

1. INTRODUCTION

Monotonicity is one of the most widely used assumptions imposed on functions explaining the relation between natural processes and has various applications in basic science, social science, and other important research areas. For instance, [23] studied monotone comparative statics, which has been a topic of interest in Economics for many years. [15] investigated the lack of monotonicity to make out certain phenomena related to the strategic behavior of economic agents, and this topic is well-known in industrial management. For various applications of monotonicity in Econometric theory, the readers are referred to [21]. In medical science, the growth curves are generally monotone. In Environmental science, it is an established fact that the number of days until the freezing of Lake Mendota can be considered as a monotone function over the years (see, e.g., [6]). Besides, the concept of monotonicity has been used in capacity reduction problems in information theory as well (see, e.g., [12]). Such monotonicity property is very often visible when the covariates are multivariate as well. As [4] pointed out that blood pressure is a monotone function of the use of tobacco and the body weight (see, e.g., [25]).

However, it may very often happen that the real data obtained for the examples in the preceding paragraph are generated from a complex stochastic model, and there is no prior guarantee that the original function is a monotone function. In fact, this situation may likely happen more when the covariate is multivariate, which is of interest in this article. In the literature, such a model is called a mis-specified model; for details on mis-specified models, see [7] and a few references therein. In this article, our theory allows the model to be mis-specified with dependent error random variables and investigates the large-sample properties of the least square isotonic regression estimator (LSEMIR) when the actual multivariate function is not componentwise isotonic. For such mis-specified multivariate isotonic regression models, we establish that the LSEMIR uniformly converges almost surely to a certain multivariate isotonic function, which is closest to the original function in a certain sense.

1.1. **Literature Review and Major Contribution.**

There are several articles in the literature regarding monotonicity constraint on univariate regression function (see, e.g., [8], [1], [3], [5], [14], [19], [16] and the references therein). However, the number of research articles on multivariate isotonic regression models is not too many. In the 1970s, [20] and [28] studied the consistency property of the least squares estimator when the number of covariates is more than

2010 *Mathematics Subject Classification.* 62G08, 62G05, 60B10.

one. To the best of our knowledge, after a gap of almost forty years, the work on multivariate isotonic regression function was revamped in the last decade. [11] studied the risk bound of the least squares estimator for bivariate isotonic regression function, and [17] also investigated a similar problem for general dimension. Recently, [4] derived the pointwise asymptotic distribution of the least squares estimator of multivariate isotonic regression function after appropriate normalization and developed a consistent test based on the least squares estimator. However, none of the above research articles considered the mis-specified model and the dependent error random variables, unlike this research article.

The main contribution of this article to the literature is the characterization of the least squares estimator of multivariate isotonic regression function when the model is mis-specified, and the errors belong to a certain stationary process. In this context, the almost sure convergence of the LSEMIR is established under mild conditions on the regression function and the dependent error random variables. The connection between the original function and the limiting function is well characterized. Apart from the aforementioned issues related to the mis-specified multivariate regression model and the dependent errors, one of the advantages of the LSEMIR is that one does not need to deal with any smoothing parameter to implement this methodology.

1.2. **Mathematical Challenges.**

To characterize the LSEMIR for mis-specified multivariate isotonic regression models with dependent errors, there are three fold mathematical challenges. The first fold is related to the domain of the covariates. For instance, when the marginal covariates are defined on $\mathbb{R}$, and the unknown regression function is a real-valued multivariate continuous function, we cannot use the projection theorem of Hilbert space to characterize the closest monotone approximation, as the space of continuous functions is not a Hilbert space. In order to have the structure of Hilbert space, one possibility is to assume that the covariates are randomly distributed as some d-dimensional ($d \geq 1$) random variable U, and $E[m^2(U)] < \infty$, where m is the unknown true regression function. However, the condition $E[m^2(U)] < \infty$ may not be satisfied for many cases. To avoid such problems, we here consider the unknown regression function is defined on $[0,1]^d$, which has one-to-one correspondence with any compact set $[a,b]^d$ for any $a \in \mathbb{R}$ and $b \in \mathbb{R}$. This structure enables us to use the projection theorem in Hilbert space, which has a key role in our theoretical study.

The second fold is related to the working formula of the least squares estimator. The min-max representation of the least squares estimator (see, e.g., [20]) is common across the literature of isotonic regression. However, as pointed out in [4], this representation is difficult to work with due to the complicated structure of the sets. In order to derive the pointwise asymptotic distribution of the least squares estimator, [4] established a geometric structure of the least squares estimator based on greatest d-convex minorant (GCM). This geometric characterization of the least squares estimator helps us to crack a major part of the proof, which may be more difficult for min-max representation of the least squares estimator.

The third fold is involved with the dependence of the error random variables. In most of the papers on isotonic regression mentioned in the earlier paragraphs, it is assumed that the error random variables are i.i.d. random variables. Inference for univariate isotonic regression under dependent errors have been studied in [2], [30] and [3]. However, to the best of our knowledge, there have not been any studies on multivariate isotonic regression under dependence. We here assume that the collection of error random variables is a weak-dependent stationary processes, and under this set up, we use modern results of the law of large numbers and the central limit theorem for such process in various places of the proofs. On the cost of these technical difficulties, such dependence structures among the error random variables widen the range of applications of the proposed methodology.

1.3. **Organization of the Article.**

The article is organized as follows. In Section 2, we describe the model and define the LSEMIR estimator along with a continuous version of the LSEMIR estimator. Section 2.2 explains the concept related to projection in Hilbert space in the context of this problem, and the ideas β-mixing stochastic process and its related issues are discussed in Section 2.3. The uniformly almost sure convergence of the LSEMIR is studied, and the characterization of the limiting function is thoroughly explored in Section 2.4. Some concluding remarks are discussed in Section 3, and at the end, all technical details are provided in Section 4.

2. Model and Main Result

2.1. Model and Its Estimator.

Consider a collection of points $\mathbf{x_i}$, where the index $\mathbf{i}$ is a d-tuple $(i_1, i_2, \ldots, i_d)$ such that $i_k = 1, \ldots, n_k$ and $x_{i_k,k} = \frac{i_k-1}{n_k-1}$. In particular, the collection of $\mathbf{x_i}$'s constitute a grid on $[0,1]^d$ of size $n_1 \times n_2 \times \ldots \times n_d$ with equidistant points. Consider data $y_{\mathbf{i}} \in \mathbb{R}$ observed on this grid following the regression model

$$y_{\mathbf{i}} = m(\mathbf{x_i}) + \epsilon_{\mathbf{i}}.$$

Equivalently, we can write the regression model as

$$y_{i_1\ldots i_d} = m(x_{i_1,1}, \ldots, x_{i_d,d}) + \epsilon_{i_1\ldots i_d}, \tag{2.1}$$

for $i_k = 1, 2, \ldots, n_k$ and $k = 1, 2, \ldots, d$. Let $n = \prod_{i=1}^d n_i$ be the total sample size. We assume that the regression function $m : [0,1]^d \mapsto \mathbb{R}$ is a continuously differentiable function, and $\epsilon_{\mathbf{i}}$s are stationary random variables with zero mean and finite variance generated from an absolutely regular random process or β mixing process. In this context, we would like to emphasize that the function m may not be co-ordinatewise non-decreasing function. Apart from the aforesaid assumptions, a few more assumptions are also required, which are stated formally before the statement of the results.

Let us now define the least squares estimator of $m(.)$ explicitly. The construction of our estimator has two steps. In the first step, we define the least squares estimator defined on the design points and extend it to a piecewise constant function, and in the second step, we construct a continuous version of the estimator on $[0,1]^d$. The formal definition is as follows :
Step 1 : Let $\hat{m}_n(.)$ be the least squares estimator of $m(.)$ on the design points, i.e.,

$$\hat{m}_n = arg \min_{m\in\mathcal{M}} \sum_{i_1=1}^{n_1} \cdots \sum_{i_d=1}^{n_d} \left(y_{i_1\ldots i_d} - m((x_{i_1,1}, \ldots, x_{i_d,d})\right)^2, \tag{2.2}$$

where $\mathcal{M}$ is the class of co-ordinate wise non-decreasing functions on $\mathbb{R}^d$. The solution of this optimization problem can be obtained by multivariate pooled adjacent violators algorithm (PAVA) described is [18], and the solution is a piece-wise constant left-continuous function. This is in fact a widely used traditional isotonic regression estimator.
Step 2 : In this step, we construct a continuous version of $\hat{m}_n$. To explicitly construct this continuous version, $\widetilde{m}_n$, we interpolate it sequentially across each coordinate. First define $\widetilde{m}_n(\mathbf{x}) = \hat{m}_n(\mathbf{x})$ if $\mathbf{x}$ is a design point, i.e., $\mathbf{x} = (x_{i_1,1}, x_{i_2,2}, \ldots, x_{i_d,d})$ where $x_{i_k,k} = (i_k-1)/(n_k-1)$ for $i_k = 1, \ldots, n_k$ and $k = 1, \ldots, d$.

Now, consider the interpolation across the first coordinate, i.e., we want to define $\widetilde{m}_n$ for $\mathbf{x} = (x_1, x_{i_2,2}, \ldots, x_{i_d,d})$ with $x_1 \in (0,1), x_{i_k,k} = \frac{i_k-1}{n_k-1}$ for $i_k = 1, \ldots, n_k$ and $k = 2, \ldots, d$. At this point, fix $\mathbf{x}$. Given this point and its first coordinate $x_1 \in (0,1)$, let x_l and x_r be the grid points on the first coordinate at the immediate left and right of x_1. In particular, let $j \in \mathbb{N}$ be such that $j/n_1 < x_1 \leq (j+1)/n_1$, then $x_{l,1} = j/n_1$ and $x_{r,1} = (j+1)/n_1$. With this, we define $\widetilde{m}_n$

$$\widetilde{m}_n(\mathbf{x}) = \widehat{m}_n(\mathbf{x}_l) + (\widehat{m}_n(\mathbf{x}_r) - \widehat{m}_n(\mathbf{x}_l))\frac{(x_1 - x_{l,1})}{(x_{r,1} - x_{l,1})}, \tag{2.3}$$

where $\mathbf{x}_l = (x_{l,1}, x_{i_2,2}, \ldots, x_{i_d,d})$ and $\mathbf{x}_r = (x_{r,1}, x_{i_2,2}, \ldots, x_{i_d,d})$.

Given we have already interpolated across coordinates $1, \dots, (d_0-1)$, consider now interpolation across coordinate $1 < d_0 \leq d$. At this step, we fix a point $\mathbf{x} = (x_1, \dots, x_{d_0}, x_{i_{d_0+1}, d_0+1}, \dots, x_{i_d, d})$, where $x_1, \dots, x_{d_0} \in (0,1)$ and $x_{i_k,k} = (i_k - 1)/(n_k - 1)$ for $i_k = 1, 2, \dots, n_k$ and $k = (d_0+1), \dots, d$. Pick x_{l,d_0} and x_{r,d_0} as the immediate left and right points on the d_0-th coordinate grid of x_{d_0}. Finally, define

$$\widetilde{m}_n(\mathbf{x}) = \widehat{m}_n(\mathbf{x}_l) + (\widehat{m}_n(\mathbf{x}_r) - \widehat{m}_n(\mathbf{x}_l)) \frac{(x_{d_0} - x_{l,d_0})}{(x_{r,d_0} - x_{l,d_0})}, \tag{2.4}$$

where we use the notations $\mathbf{x}_l = (x_1, \dots, x_{d_0-1}, x_{l,d_0}, x_{i_{d_0+1}, d_0+1}, \dots, x_{i_d,d})$, and $\mathbf{x}_r = (x_1, \dots, x_{d_0-1}, x_{r,d_0}, x_{i_{d_0+1}, d_0+1}, \dots, x_{i_d,d})$.

Remark 2.1. Note that it is immediate from the construction that $\widetilde{m}_n(.)$ is a continuous and coordinatewise non-decreasing function on $\mathbb{R}^d$. Further, $\widetilde{m}_n(\mathbf{x}_i) = \widehat{m}_n(\mathbf{x}_i)$, where $\mathbf{x}_i$ is a design point. Therefore, $\widetilde{m}_n$ is in fact a solution of (2.2).

The construction of $\widetilde{m}_n$ is primarily an artifact of our technical derivations, as the continuity of $\widetilde{m}_n$ provides more mathematical flexibility. However, this also gives a continuous version of the traditional estimator. We later show both these constructions are asymptotically equivalent (Proposition 4.1), and $\widetilde{m}_n$ does not provide any particular advantage over the traditional estimator for point-wise inference.

2.2. **Basic Concepts.** In this article, we investigate various statistical properties of the classical estimator $\hat{m}_n$, i.e., the LSEMIR defined in Step 1, and the similar results for $\widetilde{m}_n$ (proposed in Step 2) will be discussed subsequently. Before stating the theoretical results, we need to introduce a new function m^*, which is the closest monotone approximation of m in some sense under appropriate assumptions. Let us consider the Hilbert space

$$\mathcal{H} = \left\{ g : [0,1]^d \mapsto \mathbb{R}, \int\limits_{[0,1]^d} g^2(x)dx < \infty \right\}$$

with the inner product

$$\langle g, f \rangle = \int g(x)f(x)dx,$$

and the induced norm

$$\|g\|_2 = \langle g, g \rangle = \int g^2(x)dx, \text{ for all } g \in \mathcal{H}.$$

Observe that $\mathcal{M}$ (defined in Step 1 in Section 2.1) is a closed convex subset of $\mathcal{H}$ (see Lemma 4.3). Now, an application of projection theorem (see [29], p. 312–313) implies that given any function $m \in \mathcal{H}$, there exists a unique m^* in $\mathcal{M}$, such that

$$m^* = \arg\min_{f \in \mathcal{M}} \|m - f\|_2. \tag{2.5}$$

Moreover, m^* can be characterized by the following equations: (see [10], Corollary 2.3)

$$\langle m - m^*, m^* \rangle = 0, \ \& \ \langle m - m^*, g \rangle \leq 0, \text{ for all } g \in \mathcal{M}. \tag{2.6}$$

In particular, in the regression context, m^* is the element in $\mathcal{M}$, that is closest to the true regression function m in terms of the norm distance as in (2.5). Specifically, if $m \in \mathcal{H}$ is a coordinatewise non-decreasing function itself, i.e., the model is properly specified, we then have $m = m^*$.

As suggested by a reviewer, we discuss an explicit example to see how a mis-specified regression model $m \in \mathcal{H}$ but $m \notin \mathcal{M}$ and the corresponding $m^* \in \mathcal{M}$ differ from each other.
Example : The example is constructed using the idea of shifted Legendre polynomials (see, e.g., [26]) and the aforementioned projection theorem. The notations used in this example are

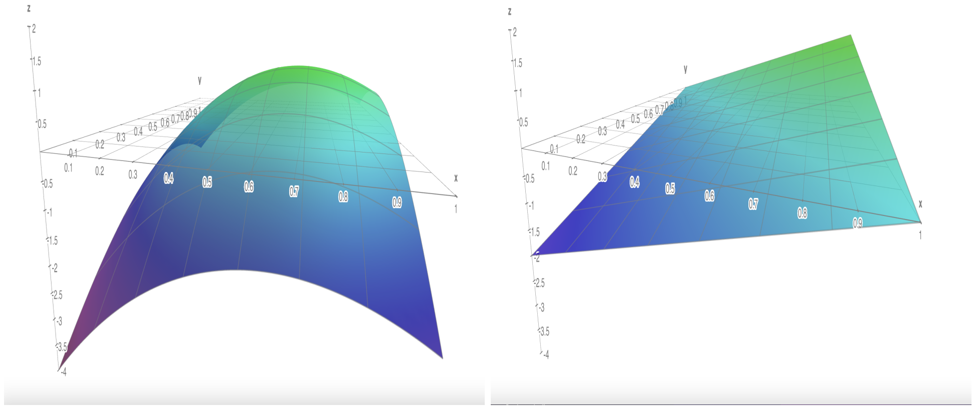

FIGURE 1. *The graphs of m and m^* for different values of $x \in (0,1)$ and $y \in (0,1)$.*

self-standing only for this example. Let us consider $m(x,y) = (2x-1) - (6x^2 - 6x + 1) + (2y-1) - (6y^2 - 6y + 1)$, where $(x,y) \in (0,1) \times (0,1)$. Note that $m \in \mathcal{H}$ but *not* co-ordinatewise non-decreasing, i.e., $m \notin \mathcal{M}$. Using the concept of Legendre polynomials, we have $m^*(x,y) = (2x-1) + (2y-1)$ since for any $x \in (0,1)$, $\langle (2x-1), -(6x^2 - 6x + 1)\rangle = 0$ and $\langle -(6x^2 - 6x + 1), g\rangle < 0$ for any $g \in \mathcal{M}$ (when $d = 1$). The similar argument holds for $d = 2$ as well. Also, note that $m^* \in \mathcal{H}$ and co-ordinatewise non-decreasing as well, i.e., $m^* \in \mathcal{M}$. For the purpose of illustration, Figure 1 plots $m(x,y)$ and $m^*(x,y)$ for $(x,y) \in (0,1) \times (0,1)$. In this context, we would like to mention that one can construct many more examples using the concept of Legendre or such orthogonal polynomials along with the fact of projection theorem.

2.3. **Dependence Structure: β-mixing.** We now want to briefly discuss the mixing stochastic process as we mentioned that the error random variables follow β-mixing stochastic process, and hence, it is needless to say that this dependence structure has noteworthy impact on our theoretical study. Strictly speaking, we consider data generated from an absolutely regular mixing stochastic process.

Definition 2.2. The β-mixing coefficients between two sigma-algebras $\mathcal{A}$ and $\mathcal{B}$ is defined as

$$\beta(\mathcal{A}, \mathcal{B}) = \sup_{\substack{A_1, A_2, \ldots A_r \in \mathcal{A} \\ B_1, B_2, \ldots, B_s \in \mathcal{B} \\ \uplus_{i=1}^r A_i = \uplus_{j=1}^s B_j = \Omega}} \frac{1}{2} \sum_{i=1}^r \sum_{j=1}^s |P(A_i \cap B_j) - P(A_i)P(B_j)|,$$

where Ω is the sample space, $A_1, \ldots, A_r$ are r many arbitrary events in $\mathcal{A}$, and $B_1, \ldots, B_s$ are s many arbitrary events in $\mathcal{B}$.

For a collection of random variables $\{\epsilon_{i_1 \ldots i_d}\}$ indexed by $i_k \in \mathbb{Z}$, $k = 1, 2, \ldots, d$, define the β-mixing coefficients as

(2.7)

$$\beta_{k_1, \ldots, k_d} = \sup_{(l_1, \ldots, l_d) \in \mathbb{Z}^d} \beta(\sigma(\{\epsilon_{i_1 \ldots i_d} : i_j \leq l_j - k_j, j = 1, \ldots, d\}), \sigma(\{\epsilon_{i_1 \ldots i_d} : i_j \geq l_j, j = 1, \ldots, d\})),$$

and the sequence is said to be β-mixing or absolutely regular mixing sequence if $\beta_{k_1, \ldots, k_d} \to 0$ as $k_j \to \infty$ for at least one $j = 1, 2, \ldots, d$. Here $\sigma(.)$ denotes the smallest σ-field generated by the collection of random variables inside the parenthesis. Observe that $\beta_{k_1, \ldots, k_d}$ measures the dependence in the error sequence at lag k_j at the j-th coordinate. As we know, for most of the natural processes, the nearer observations are more dependent than the further ones, and therefore, $\beta_{k_1,\dots,k_d}$ is expected to decrease as $k_j \to \infty$ for any j. Moreover, if the sequence $\beta_{k_1,k_2,\dots,k_d}$ is absolutely summable, one can show that Var $\left(\frac{1}{\sqrt{n}}\sum_{i_1=1}^{n_1}\cdots\sum_{i_d=1}^{n_d}\epsilon_{i_1\dots i_d}\right) < \infty$, and the sequence $\{\epsilon_{i_1\dots i_d}\}$ exhibits weak-dependence or short-memory dependence (see [27]).

The assumption of β-mixing is quite common in the literature. [13] showed that Markov chains under Harris recurrence conditions are geometrically β-mixing. For example, if $(X_t,\sigma_t)_{t\in\mathbb{Z}}$ is a stationary GARCH process with appropriate parameters such that the distribution of noise sequence $\{\epsilon\}_{i\geq 0}$ is absolutely continuous with Lebesgue measure being strictly positive in a neighborhood of zero and $\mathbb{E}|\epsilon_0|^s < \infty$ for some $s > 0$, then both the sequences $(X_t)_{t\in\mathbb{Z}}$ and $(X_t^2)_{t\in\mathbb{Z}}$ are geometrically β-mixing (see [9]). [24] showed that stationary vector valued ARMA processes with innovations from an absolutely continuous distribution with respect to Lebesgue measure are geometrically β-mixing. These various applications of β-mixing stochastic processes motivated us to study this problem with dependent errors, which can be defined through a certain β-mixing.

2.4. **Almost sure Convergence : Characterization.**

We start by stating a few assumptions that will be required to establish the results related to the almost sure convergence of $\hat{m}_n$.
(C1) The true regression function $m : [0,1]^d \to \mathbb{R}$ is a continuously differentiable function.
(C2) Let $\{\epsilon_{i_1\dots i_d}\}$ be second-order stationary across each coordinate with zero mean and β-mixing coefficients $\beta_{n_1,\dots,n_d}$ satisfying either one of the following two conditions:

(a) (geometrical β-mixing) $E(\epsilon^2_{i_1\dots i_d}) < \infty$ and $\beta_{n_1\dots n_d} = \prod_{j=1}^d O(a^{n_j})$ for some $0 < a < 1$.

(b) (polynomial β-mixing) $E(|\epsilon_{i_1\dots i_d}|^{2r}) < \infty$ for some $r > 1$ and $\beta_{n_1\dots n_d} = \prod_{j=1}^d O(n_j^{-L})$ for some $L > r/(r-1)$.

Remark 2.3. Note that condition (C1) is trivially true for many choices of m in practice. Condition (C2) needs attention since there is a trade-off between the existence of the moments of the error random variables and the strength of the dependence. In particular, the existence of higher moments allows stronger dependence among the errors. Observe that the moment conditions stated in (a) and (b) will hold for many distributions, such as Gaussian and Laplace distributions. To be summarized, the conditions are realistic; in particular, (C2) ensures that one does not need even i.i.d. sequence of error random variables to implement this methodology, which broadens the applicability of the methodology.

We now state the result on uniformly almost sure convergence of $\hat{m}_n$, which is one of the major characterizations of $\hat{m}_n$.

Theorem 2.4. *Assume that $\widehat{m}_n$ is the solution of* (2.2)*, and m^* is the same as defined in* (2.5)*. Under (C1) and (C2), we have*

$$\sup_{\mathbf{x}\in(0,1)^d} |\widehat{m}_n(\mathbf{x}) - m^*(\mathbf{x})| \to 0, \text{ almost surely as } n\to\infty. \tag{2.8}$$

The assertion of Theorem 2.4 implies that the LSEMIR, i.e., $\hat{m}_n$ uniformly converges almost surely to a function, which is a co-ordinatewise non-decreasing function and the closest to the true function m (see (2.5)) as the true function $m \in \mathcal{H}$ (see Lemma 4.2). In this context, we would like to recall the alternative continuous version of the classical estimator, i.e., $\widetilde{m}_n$ introduced in Step 2 earlier in this section. In the following proposition, we state the result related to $\widetilde{m}_n$ as well.

Proposition 2.1. *Let $\widetilde{m}_n$ be the same as defined in* (2.4)*, and m^* is the same as defined in* (2.5)*. Then for any $0 < a < b < 1$, under (C1)–(C2), we have*

$$\sup_{\mathbf{x}\in[a,b]^d} |\widetilde{m}_n(\mathbf{x}) - m^*(\mathbf{x})| \to 0, \text{ almost surely as } n\to\infty. \tag{2.9}$$

Proposition 2.1 asserts that $\widetilde{m}_n$ has the same characterization property as the classical estimator $\hat{m}_n$.

Remark 2.5. The results in Theorem 2.4 and Proposition 2.1 are established for the design points defined on a regular grid. If the design points are random, such consistency results can be extended using similar techniques. However, for random design points, the limiting regression function won't be the same as m^* obtained for the equidistant design points. It is expected that the limiting regression function depends on the distribution of the design points when the design points are random.

3. Concluding Remarks

This article studies the least squares estimator of the multivariate isotonic regression function when the model is mis-specified, and the errors are dependent random variables. Under mild conditions, we observe that the least squares estimator converges uniformly to a certain monotone function, which is closest to the original function in some sense. Apart from the issues related to the large sample properties, the performance of the least squares estimator is also investigated for various models when the sample size is finite.

The "mis-specified" or the "wrong" model perspective has important implications for practice since strictly speaking, any model cannot be a "correct" model for a given data, and it is needless to mention that this view is one of the major motivator of creating the research field like model selection. However, the literature has not paid much attention to the mis-specified model, and the reason lies in the fact related to technical difficulties. The study on mis-specified models will be more available in the literature once the associated technical difficulties are overcome to a large extent.

Acknowledgement

The authors are thankful to the Editor-in-Chief Prof. Yuliya Mishura, an anonymous associate editor and two anonymous reviewers for their many useful comments and suggestions, which improve the article significantly. Subhra Sankar Dhar is supported by CRG (File no : CRG/2022/001489) and MATRICS (File No : MTR/2019/000039), research grants from the SERB, Government of India.

4. Technical Details and Proofs

4.1. **Geometric Characterization of Isotonic Regression Estimator.** We begin with briefly reviewing the geometric characterization of the isotonic regression estimator, which will be extensively used to prove our main results. A detailed description can be found in Section 2.1 of [4].

Let us start with the concept of the left-slope of a multivariate function. For a real-valued function G defined on $\mathbb{R}^d$, let $\partial_\ell G(\mathbf{x})$ denote the mixed left partial derivative with respect to $\mathbf{x} \in \mathbb{R}^d$. More precisely, if $\mathbf{x} = (x_1, x_2, \ldots, x_d)$, we have

$$\partial_\ell G(\mathbf{x}) = \partial_\ell G(x_1, \ldots, x_d) = \frac{\partial^d G(x_1-, \ldots, x_d-)}{\partial x_1 \ldots \partial x_d},$$

where $\partial f(x_{0,1}, \ldots, x_{0,d})/\partial x_k$ denotes the partial derivative of f with respect to k-th coordinate x_k at the point $\mathbf{x}_0 = (x_{0,1}, \ldots, x_{0,d})$. The right-slope $\partial_r G$ can be defined similarly.

Definition 4.1. We define the class of d-convex functions on $I \subset \mathbb{R}^d$ to be

$$\mathcal{C}_I := \{G : I \mapsto \mathbb{R}, G \text{ is convex on } I \text{ and } \partial_\ell G \text{ is coordinate-wise non-decreasing}\}. \tag{4.1}$$

For any real-valued function S defined on $I \subset \mathbb{R}^d$, we define d-GCM $T_I(S)$ of S as the pointwise supremum of all d-convex function which lies below S, i.e.,

$$T_I(S)(\mathbf{x}) = \sup_{G \in \mathcal{C}_I; G \leq S} G(\mathbf{x}), \ \mathbf{x} \in I. \tag{4.2}$$

For the sake of notational simplicity, if $I = \mathbb{R}^d$, we drop the subscript and write $T_{\mathbb{R}^d}(S)$ as $T(S)$.

We denote the cumulative sum diagram of the data as $\mathbb{S}_n$. To be precise, we define $\mathbb{S}_n$ on $[0,1]^d$ on the design points $\mathbf{x}_i = (x_{1,i_1}, \ldots, x_{d,i_d})$ as

$$\mathbb{S}_n(\mathbf{x}_i) = \mathbb{S}_n(x_{1,i_1}, \ldots, x_{d,i_d}) = \frac{1}{n_1 \ldots n_d} \sum_{l_1 \leq i_1} \cdots \sum_{l_d \leq i_d} y_{l_1 \ldots l_d}, \text{ for } i_k = 0, \ldots, n_k; k = 1, \ldots, d$$

with the notations $x_{k,0} = 0$ and $y_{i_1 \ldots i_d} = 0$ if $i_k = 0$ for any k. The process $\mathbb{S}_n$ is then interpolated linearly at each coordinate in between the design points, and this interpolation is done sequentially for each coordinate. For details on the construction of S_n, the readers are referred to Section 2 of [4].

Finally, Theorem 2 of [4] asserts that

$$\widehat{m}_n(\mathbf{x}) = \partial_l T_{[0,1]^d}(\mathbb{S}_n)(\mathbf{x}), \text{ for all } \mathbf{x} \in [0,1]^d. \tag{4.3}$$

4.2. **Proof of Theorem 2.4.**

4.2.1. *Outline of the Proof of Theorem 2.4.* We start by noting that the projection equations (2.6) can be restated as

$$\langle m - m^*, g - m^* \rangle \leq 0, \text{ for all } g \in \mathcal{M}, \tag{4.4}$$

where $\mathcal{M}$ is same as the defined in (2.2). Firstly, we will consider the continuous version of the estimator $\widetilde{m}_n$ defined in (2.4). In order to show the uniform convergence in (2.8), we will show the uniform convergence of $\widetilde{m}_n$ to m^* (Proposition 2.1). Secondly, we will establish the almost sure uniform convergence of $\widehat{m}_n - \widetilde{m}_n$ to 0 as $n \to \infty$. Finally, an application of triangle inequality upon above said two results will complete the proof.

4.2.2. *Proof of Proposition 2.1.* For notational convenience denote $J_d = \{(i_1, \ldots, i_d) \in \mathbb{N}^d : 1 \leq i_j \leq n_j, j = 1, 2, \ldots, d\}$. Note that the cardinality of J_d is n. Step 1: First observe that the quantity

$$S_n(f) = \frac{1}{n} \sum_{\mathbf{i} \in J_d} (y_{\mathbf{i}} - m(\mathbf{x}_{\mathbf{i}}))^2$$

is minimized at $f = \widetilde{m}_n$. Therefore, one can write

$$\begin{aligned} S_n(m^*) \geq S_n(\widetilde{m}_n) =& \frac{1}{n} \sum_{\mathbf{i} \in J_d} (y_{\mathbf{i}} - m^*(\mathbf{x}_{\mathbf{i}}) + m^*(\mathbf{x}_{\mathbf{i}}) - \widetilde{m}_n(\mathbf{x}_{\mathbf{i}}))^2 \\ =& S_n(m^*) + \frac{2}{n} \sum_{\mathbf{i} \in J_d} (y_{\mathbf{i}} - m^*(\mathbf{x}_{\mathbf{i}}))(m^*(\mathbf{x}_{\mathbf{i}}) - \widetilde{m}_n(\mathbf{x}_{\mathbf{i}})) \\ &+ \frac{1}{n} \sum_{\mathbf{i} \in J_d} (\widetilde{m}_n(\mathbf{x}_{\mathbf{i}}) - m^*(\mathbf{x}_{\mathbf{i}}))^2. \end{aligned}$$

Cancelling $S_n(m^*)$ from both sides and rearranging the terms, we have

$$\frac{1}{n} \sum_{\mathbf{i} \in J_d} (\widetilde{m}_n(\mathbf{x}_{\mathbf{i}}) - m^*(\mathbf{x}_{\mathbf{i}}))^2 \leq \frac{2}{n} \sum_{\mathbf{i} \in J_d} (y_{\mathbf{i}} - m^*(\mathbf{x}_{\mathbf{i}}))(\widetilde{m}_n(\mathbf{x}_{\mathbf{i}}) - m^*(\mathbf{x}_{\mathbf{i}})). \tag{4.5}$$

We would now like to point out that the right hand side of (4.5) can be interpreted as the sample version of the quantity $\langle m - m^*, \tilde{m} - m^* \rangle$ as $y_{\mathbf{i}}$ is $m(\mathbf{x}_{\mathbf{i}})$ with some additive noise. The quantity

$\langle m - m^*, \tilde{m} - m^* \rangle$ is negative by (2.6). However, due to randomness involved through the regression errors in this term, one needs to have some further technicalities to conclude that the sum in (4.5) is asymptotically bounded above by 0.

Step 2: We now claim that $\widetilde{m}_n$ is uniformly bounded on $[a, b]^d$ for any $0 < a < b < 1$ almost surely.
To see it, note that by monotonicity of $\widetilde{m}_n$, we have

$$\limsup_n \sup_{\mathbf{x}\in[a,b]^d} |\widetilde{m}_n(\mathbf{x})| \leq \limsup_n \max\left\{|\widetilde{m}_n(\mathbf{b})|, |\widetilde{m}_n(\mathbf{a})|\right\}, \tag{4.6}$$

where $\mathbf{a} = (a, \ldots, a) \in \mathbb{R}^d$ and $\mathbf{b} = (b, \ldots, b) \in \mathbb{R}^d$. Due to the grid structure of the design points, given $b < 1$, there exists a design point $\mathbf{x}_r$ such that $\mathbf{b} < \mathbf{x}_r < \mathbf{1}$ for large enough n, where $\mathbf{1} = (1, \ldots, 1) \in \mathbb{R}^d$, and the inequality is the componentwise inequality. Without loss of generality, assume $\widetilde{m}_n(\mathbf{b}) > 0$. Now fix $M > 0$, and using (4.3), one can write

$$\begin{aligned}
\mathbb{P}\big(\widetilde{m}_n(\mathbf{b}) > M\big) &\leq \mathbb{P}\big(\widetilde{m}_n(\mathbf{x}_r) > M\big)\\
&= \mathbb{P}\big(\partial_l T_{[0,1]^d}(\mathbb{S}_n)(\mathbf{x}_r) > M\big)\\
&= \mathbb{P}\big(T_{[0,1]^d}(\mathbb{S}_n)(\mathbf{x}) > l, \text{ for all } \mathbf{x} > \mathbf{x}_r \text{ componentwise}\big)
\end{aligned}$$

where l is a linear function with slope M and passing through $T_{[0,1]^d}(\mathbb{S}_n)(\mathbf{x}_r)$ at $\mathbf{x}_r$. The last quantity implies

$$\begin{aligned}
\mathbb{P}\big(\widetilde{m}_n(\mathbf{b}) > M\big) &\leq \mathbb{P}\left(\mathbb{S}_n(\mathbf{x}) > T_{[0,1]^d}(\mathbb{S}_n)(\mathbf{x}_r) + M\prod_{k=1}^{d}(x_k - x_{r,k}), \text{ for all } \mathbf{1} \geq \mathbf{x} > \mathbf{x}_r \text{ component wise}\right)\\
&\leq \mathbb{P}\left(\mathbb{S}_n(\mathbf{x}_{r+1}) - \mathbb{S}(\mathbf{x}_r) > T_{[0,1]^d}(\mathbb{S}_n)(\mathbf{x}_r) - \mathbb{S}_n(\mathbf{x}_r) + M\prod_{k=1}^{d}(x_{r+1,k} - x_{r,k})\right)\\
&\leq \mathbb{P}(Y_{r+1} > M),
\end{aligned}$$

and the last quantity tends to 0 as $M \to \infty$. Arguing in a similar way, one can establish that $\widetilde{m}_n(\mathbf{a})$ is also stochastically bounded. Using (4.6) along with this fact gives us

$$\limsup_n \sup_{\mathbf{x}\in[a,b]^d} |\widetilde{m}_n(\mathbf{x})| \leq X(\mathbf{a}, \mathbf{b}), \tag{4.7}$$

where $X(\mathbf{a}, \mathbf{b})$ is a valid random variable independent of n and depends on $\mathbf{a} = (a, \ldots, a) \in \mathbb{R}^d$ and $\mathbf{b} = (b, \ldots, b) \in \mathbb{R}^d$. To this point, choose an $\epsilon > 0$, and choose $c(\mathbf{a}, \mathbf{b}, \epsilon)$ such that $P(X(\mathbf{a}, \mathbf{b}) > c(\mathbf{a}, \mathbf{b}, \epsilon)) < \epsilon$. Thus, we have

$$P\left(\{\omega : \limsup_n \sup_{\mathbf{x}\in[a,b]^d} |\widetilde{m}_n(\mathbf{x})|(\omega) < c(\mathbf{a}, \mathbf{b}, \epsilon)\}\right) > 1 - \epsilon.$$

To this end, define

$$\Omega_0 = \{\omega : \limsup_n \sup_{\mathbf{x}\in[a,b]^d} |\widetilde{m}_n(\mathbf{x})|(\omega) < c(\mathbf{a}, \mathbf{b}, \epsilon)\}.$$

Step 3: Next, it follows from Lemma 4.5 that $\widetilde{m}_n$ is uniformly bounded on $[a, b]^d$ and piece-wise linear, which implies that $\widetilde{m}_n$ is Lipschitz over $[a, b]^d$ uniformly in n on Ω_0, a set of probability at least $1 - \epsilon$, i.e., on this set,

$$|\widetilde{m}_n(\mathbf{x}) - \widetilde{m}_n(\mathbf{y})| \leq 2c(\mathbf{a}, \mathbf{b}, \epsilon)\|\mathbf{x} - \mathbf{y}\|,$$

for $\mathbf{x}, \mathbf{y} \in [a, b]^d$ when n is sufficiently large.

Consider now the class of functions

$$\mathcal{C} := \left\{ \begin{array}{l} h : [a,b]^d \mapsto \mathbb{R} : h \text{ is co-ordinate wise non-decreasing,} \\ |h(\mathbf{x})| \le c(\mathbf{a},\mathbf{b},\epsilon), |h(\mathbf{x}) - h(\mathbf{y})| \le 2c(\mathbf{a},\mathbf{b},\epsilon)\|\mathbf{x}-\mathbf{y}\| \end{array} \right\}.$$

Observe that on Ω_0, $\widetilde{m}_n \in \mathcal{C}$ for sufficiently large n, and $\mathcal{C}$ is compact with respect to the norm

$$d(h_1,h_2) = \sup_{\mathbf{x}\in[a,b]^d} |h_1(\mathbf{x}) - h_2(\mathbf{x})|.$$

In view of the fact that $\mathcal{C}$ is compact, given $\eta > 0$, one can find $h_1, h_2, \ldots, h_m \in \mathcal{C}$ such that

$$\bigcup_{i=1}^{m} \left\{ h \in \mathcal{C} : d(h,h_i) < \eta \right\} \supset \mathcal{C},$$

and therefore on Ω_0, given $\widetilde{m}_n$, one can pick $h_J \in \mathcal{C}$ such that

$$\sup_{\mathbf{x}\in[a,b]^d} |\widetilde{m}_n(\mathbf{x}) - h_J(\mathbf{x})| = d(\widetilde{m}_n, h_J) < \eta,$$

for sufficiently large n.

Let $[a,b]^d$ be large enough such that number of design points in $[0,1]^d - [a,b]^d$ is $\le n^{1-\delta}$ for some $\delta > 0$, which follows from the condition (C3). Recall the right hand side of (4.5) and write

$$\begin{aligned} &\frac{1}{n}\sum_{\mathbf{i}\in J_d}(y_{\mathbf{i}} - m^*(\mathbf{x_i}))(\widetilde{m}_n(\mathbf{x_i}) - m^*(\mathbf{x_i})) \\ &= \frac{1}{n}\sum_{\mathbf{i}\in J_d}(y_{\mathbf{i}} - m^*(\mathbf{x_i}))(\widetilde{m}_n(\mathbf{x_i}) - h_J(\mathbf{x_i})) + \frac{1}{n}\sum_{\mathbf{i}\in J_d}(y_{\mathbf{i}} - m^*(\mathbf{x_i}))(h_J(\mathbf{x_i}) - m^*(\mathbf{x_i})) \\ &\le \frac{1}{n}\sum_{\mathbf{i}\in J_d}(y_{\mathbf{i}} - m^*(\mathbf{x_i}))(\widetilde{m}_n(\mathbf{x_i}) - h_J(\mathbf{x_i}))I(\mathbf{x_i} \in [a,b]^d) + \frac{1}{n^\delta} \\ &\quad + \frac{1}{n}\sum_{\mathbf{i}\in J_d}(m(\mathbf{x_i}) - m^*(\mathbf{x_i}))(h(\mathbf{x_i}) - m^*(\mathbf{x_i})) + \frac{1}{n}\sum_{\mathbf{i}\in J_d}\epsilon_{\mathbf{i}}(h_J(\mathbf{x_i}) - m^*(\mathbf{x_i})) \\ &= (I) + (II) + (III) + (IV). \end{aligned}$$

Since $y_{\mathbf{i}} = m(\mathbf{x_i}) + \epsilon_{\mathbf{i}}$, observe that (I) is bounded above by

$$\sup_{\mathbf{x}\in[a,b]^d} |\widetilde{m}_n(\mathbf{x_i}) - h_J(\mathbf{x_i})| \left[\frac{1}{n}\sum_{\mathbf{i}\in J_d}(m(\mathbf{x_i}) - m^*(\mathbf{x_i})) + \frac{1}{n}\sum_{\mathbf{i}\in J_d}\epsilon_{\mathbf{i}} \right].$$

Now, using the law of large numbers for strong mixing process (see [22]), we have

$$\frac{1}{n}\sum_{\mathbf{i}\in J_d}\epsilon_{\mathbf{i}} \overset{a.s.}{\to} E(\epsilon) = 0,$$

and using (C3), we have

$$\frac{1}{n}\sum_{\mathbf{i}\in J_d}(m(\mathbf{x_i}) - m^*(\mathbf{x_i})) \to \int_{[0,1]^d} (m(\mathbf{x}) - m^*(\mathbf{x}))d\mathbf{x},$$

as $n \to \infty$. Therefore, for sufficiently large n, on Ω_0, (I) is smaller than $\eta \int (m(\mathbf{x}) - m^*(\mathbf{x}))d\mathbf{x}$ since it is already established that $\sup_{\mathbf{x}\in[a,b]^d} |\widetilde{m}_n(\mathbf{x}) - h_J(\mathbf{x})| = d(\widetilde{m}_n, h_J) < \eta$ for a given $\eta > 0$ on that set.

Observe that (IV) converges to 0, as

$$E\left(\frac{1}{n}\sum_{\mathbf{i}\in J_d}\epsilon_i(h_J(\mathbf{x_i}) - m^*(\mathbf{x_i})) \right) = \frac{1}{n}\sum_{\mathbf{i}\in J_d}^{n}(h_J(\mathbf{x_i}) - m^*(\mathbf{x_i}))E(\epsilon_{\mathbf{i}}) = 0,$$

and the random variables $\epsilon_{\mathbf{i}}(h_J(\mathbf{x_i})-m^*(\mathbf{x_i}))$ are also β-mixing random variables with β-mixing coefficients satisfying (C2) (Lemma III.1 [3]). Therefore, again by strong law of large numbers for mixing process, we have

$$\frac{1}{n}\sum_{\mathbf{i}\in J_d}\epsilon_{\mathbf{i}}(h_J(\mathbf{x_i})-m^*(\mathbf{x_i}))\overset{a.s.}{\to}0$$

as $n\to\infty$. Thus, $\frac{1}{n}\sum\limits_{\mathbf{i}\in J_d}\epsilon_{\mathbf{i}}(h_J(\mathbf{x_i})-m^*(\mathbf{x_i}))<\eta^*$ for any given $\eta^*>0$ for sufficiently large n.

Next, note that (III) converges to $\langle m-m^*, h_J-m^*\rangle$, and hence, we have

$$\frac{1}{n}\sum_{\mathbf{i}\in J_d}(Y_{\mathbf{i}}-m^*(x_{\mathbf{i}}))(\widetilde{m}_n(x_{\mathbf{i}})-m^*(x_{\mathbf{i}}))\leq\langle m-m^*, h_J-m^*\rangle+\eta^{**},$$

on Ω_0, where $\eta^{**}>0$ is arbitrary. Since ϵ is arbitrary, the last inequality holds almost surely. Moreover, as $\eta^{**}>0$ is arbitrary, and $\langle m-m^*, h_J-m^*\rangle$ is negative by (4.4), the right hand side of (4.5) is negative for sufficiently large n.

Step 4: Observe that Step 1 and Step 3, which proves that the right hand side of (4.5) is negative for n sufficiently large, together implies that

$$\limsup_{n\to\infty}\frac{1}{n}\sum_{\mathbf{i}\in J_d}\left(\widetilde{m}_n(\mathbf{x_i})-m^*(\mathbf{x_i})\right)^2=0 \text{ almost surely.} \tag{4.8}$$

Step 5: Note that $[a,b]^d$ is a compact set, and let $\Lambda_1,\dots,\Lambda_p$ be an a_n-net of the set $[a,b]^d$. Using the assumption (C3), one can argue that for all $j=1,2,\dots,p$, there exists a design point $x_i\in\Lambda_j$. Further, note that m^* is componentwise increasing, and using (4.8), it is also uniformly bounded over $[a,b]^d$, and therefore, m^* is also Lipschitz.
Now, for $\mathbf{x},\mathbf{x_i}\in\Lambda_j$, we have

$$|\widetilde{m}_n(\mathbf{x})-m^*(\mathbf{x})|\leq|\widetilde{m}_n(\mathbf{x})-\widetilde{m}_n(\mathbf{x_i})|+|\widetilde{m}_n(\mathbf{x_i})-m^*(\mathbf{x_i})|+|m^*(\mathbf{x_i})-m^*(\mathbf{x})|$$

The first term and the third term can be made small due to Lipschitz continuity (in an almost sure sense for the first term), and the second term is bounded above by

$$\sup_{\mathbf{i}}|\widetilde{m}_n(\mathbf{x_i})-m^*(\mathbf{x_i})|$$

which converges to 0 by (4.8) and the fact that the number of Λ_j's are finite, one can conclude that

$$\sup_{\mathbf{x}\in[a,b]^d}|\widetilde{m}_n(\mathbf{x})-m^*(\mathbf{x})|\to 0,\ \text{ almost surely as } n\to\infty.$$

As $0<a<b<1$ are arbitrary, it proves (2.9).

4.2.3. *Proof of Uniform Convergence of* $\widehat{m}_n-\widetilde{m}_n$.

Proposition 4.1. *Let* $\widehat{m}_n$ *be the isotonic regression estimator defined in* (2.2)*, and suppose that* $\widetilde{m}_n$ *is the same as defined in* (2.4)*. Given* $\eta^{***}>0$ *and* $0<a<b<1$*, we have*

$$\sup_{\mathbf{x}\in[a,b]^d}|\widehat{m}_n(\mathbf{x})-\widetilde{m}_n(\mathbf{x})|<\eta^{***} \tag{4.9}$$

for sufficiently large n.

Proof. Note that by the construction, we have $\widetilde{m}_n(\mathbf{x})\geq\widehat{m}_n(\mathbf{x})$ for $\mathbf{x}\in[0,1]^d$ and $\widetilde{m}_n(\mathbf{x}_i)=\widehat{m}_n(\mathbf{x}_i)$ at the design points $\mathbf{x}_i$. Given $\mathbf{x}\in[a,b]^d$, let $x_{l,k}$ and $x_{r,k}$ be the immediate left and right grid point of x_k, the k-th coordinate of $\mathbf{x}$. We write $\mathbf{x_l}=(x_{l,1},\dots,x_{l,d})$ and $\mathbf{x_r}=$

$(x_{r,1}, \ldots, x_{r,d})$. Therefore $\mathbf{x_l} < \mathbf{x} \leq \mathbf{x_r}$ componentwise, and by the componentwise monotonicity of both $\widetilde{m}_n$ and $\widehat{m}_n$, we have

$$\begin{aligned}|\widehat{m}_n(\mathbf{x}) - \widetilde{m}_n(\mathbf{x})| =& \widetilde{m}_n(\mathbf{x}) - \widehat{m}_n(\mathbf{x})\\ &\leq \widetilde{m}_n(\mathbf{x_r}) - \widehat{m}_n(\mathbf{x_l})\\ &\leq \widetilde{m}_n(\mathbf{x_r}) - \widetilde{m}_n(\mathbf{x_l}).\end{aligned}$$

Therefore,

$$\sup_{\mathbf{x}\in[a,b]^d} |\widehat{m}_n(\mathbf{x}) - \widetilde{m}_n(\mathbf{x})| \leq \sup_i |\widetilde{m}_n(\mathbf{x_r}) - \widetilde{m}_n(\mathbf{x_l})|.$$

As $\widetilde{m}_n$ is Lipschitz as established in Step 3, the last quantity is bounded above by $c(\mathbf{a}, \mathbf{b})\|\mathbf{x_r} - \mathbf{x_l}\| = O(1/n)$ using (C3), and it proves the desired result. □

4.2.4. *Final Part : Proof of Theorem 2.4.*

Proof. An application of triangle inequality on the assertions of Proposition 2.1 and Proposition 4.1 gives us, for any $\eta^{****} > 0$ and $0 < a < b < 1$,

$$\sup_{\mathbf{x}\in[a,b]^d} |\widehat{m}_n(\mathbf{x}) - m^*(\mathbf{x})| \leq \eta^{****},$$

almost surely for sufficiently large n. As η^{****}, a and b are arbitrary, it proves Theorem 2.4. □

4.3. **Other Auxiliary Results.**

Lemma 4.2. *$m \in \mathcal{H}$, where $\mathcal{H}$ is the same as defined in Section 2.2.*

Proof. From the condition (C1), note that $m : [0,1]^d \to \mathbb{R}$ is a continuously differentiable function, and hence, m is uniformly bounded over $[0,1]^d$. This implies that $\int\limits_{[0,1]^d} m^2(x)dx < \infty$, and hence, $m \in \mathcal{H}$. □

Lemma 4.3. *$\mathcal{M}$ is a closed convex subset of $\mathcal{H}$, where $\mathcal{M}$ is the same as defined in Section 2, and $\mathcal{H}$ is the same as defined in Section 2.2.*

Proof. To avoid notational complexity, the proof is given for $d = 1$. The arguments will be the same for general d.

Let $\{f_n\} \in \mathcal{M}$ be a sequence of functions such that $f_n \to f$ as $n \to \infty$ in $||.||_2$ norm, which is defined in Section 2.2. Observe that since $f_n \in \mathcal{M}$, we have $\langle f_n(x) - f_n(y), x - y\rangle \geq 0$ for all x and y. Since $\langle ., . \rangle$ is a continuous function with respect to $||.||_2$ norm, we have $\langle f_n(x) - f_n(y), x - y\rangle \to \langle f(x) - f(y), x - y\rangle$ as $n \to \infty$ in $||.||_2$ norm for all x and y. Hence, $\langle f(x) - f(y), x - y\rangle \geq 0$ for all x and y, i.e., $f \in \mathcal{M}$, which implies that $\mathcal{M}$ is a closed set of $\mathcal{H}$.

Next, to establish that $\mathcal{M}$ is a convex set of $\mathcal{H}$, let us consider two arbitrary functions $f_1 \in \mathcal{M}$ and $f_2 \in \mathcal{M}$. Suppose that $t \in [0, 1]$ is an arbitrary constant. Observe that for any $x \leq y$, we have $f_1(x) \leq f_1(y)$ as $f_1 \in \mathcal{M}$, and for any $t \in [0, 1]$, we have $tf_1(x) \leq tf_1(y)$. Similarly, $(1 - t)f_2(x) \leq (1 - t)f_2(y)$ also as $f_2 \in \mathcal{M}$ and $(1 - t) \in [0, 1]$. These two facts imply that for any $t \in [0, 1]$, $tf_1(x) + (1 - t)f_2(x) \leq tf_1(y) + (1 - t)f_2(y)$ for all $x \leq y$ and any $t \in [0, 1]$. Hence, for $t \in [0, 1]$, we have $tf_1 + (1 - t)f_2 \in \mathcal{M}$, i.e., $\mathcal{M}$ is a convex set as well. The proof is now complete since it was earlier proved that $\mathcal{M}$ is a closed set of $\mathcal{H}$.

□

Lemma 4.4. *Let L be a linear function on $\mathbb{R}^d$, for any continuous function $S : \mathbb{R}^d \mapsto \mathbb{R}$, we have $T_I(S + L) = T_I(S) + L$ on I, for any interval $I \subset \mathbb{R}^d$.*

Proof. First note that $T_I(S)+L$ is a d-convex function, and $(T_I(S)+L)(\mathbf{x}) \le (S+L)(\mathbf{x})$ for all $\mathbf{x} \in I$. Therefore, $T_I(S)+L$ is a d-convex minorant of $S+L$, and by the definition of d-GCM, we have

$$T_I(S+L)(\mathbf{x}) \ge T_I(S)(\mathbf{x}) + L(\mathbf{x}) \text{ for all } \mathbf{x} \in I.$$

On the other hand, we have $(T_I(S+L)-L)(\mathbf{x}) \le S(\mathbf{x})$ for all $\mathbf{x} \in I$. Moreover, observe that $T_I(S+L)-L$ is a d-convex function itself, and therefore, $T_I(S+L)-L \le T_I(S)$ by the definition of d-GCM, and hence, the result is proved. □

Lemma 4.5. *Suppose that we have a collection of closed polyhedrons $\{\Omega_i \subset \mathbb{R}^d\}_{i=1}^k$ such that $\Omega = \cup_{i=1}^d \Omega_i$ is convex. Let f_i be linear functions on Ω_i, i.e., $f_i(\mathbf{x}) = A_i\mathbf{x} + c_i$ for $\mathbf{x} \in \Omega_i, i = 1, \ldots, k$. Moreover, assume that for any $i \neq j$, and $\mathbf{x} \in \Omega_i \cap \Omega_j$, we have $f_i(\mathbf{x}) = f_j(\mathbf{x})$. Let f be defined on Ω such that*

$$f(\mathbf{x}) = f_i(\mathbf{x}) \quad \text{if } \mathbf{x} \in \Omega_i.$$

Given arbitrary norms $\|\cdot\|_\alpha, \|\cdot\|_\beta$, the functionf is $\max_{i\le d}\{\|A_i\|_{\alpha,\beta}\}$-Lipschitz continuous with respect to $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$, i.e., for all $x_1, x_2 \in \Omega$,

$$\|f(\mathbf{x}_2) - f(\mathbf{x}_1)\|_\alpha \le \max_{i\le d}\{\|A_i\|_{\alpha,\beta}\}\|\mathbf{x}_2 - \mathbf{x}_1\|_\beta, \tag{4.10}$$

where $\|A_i\|_{\alpha,\beta} = \max_{\|\mathbf{x}\|_\beta \le 1} \|A_i\mathbf{x}\|_\alpha$.

Proof. Let $g(t) = \mathbf{x}_1 + t(\mathbf{x}_2 - \mathbf{x}_1)$, $L = \max_{i\le d}\{\|A_i\|_{\alpha,\beta}\}\|\mathbf{x}_2 - \mathbf{x}_1\|_\beta$, and $h = f \circ g$. One can see the inequality (4.5) is equivalent to

$$\|h(1) - h(0)\|_\alpha \le L. \tag{4.11}$$

Using the property that f is well-defined on $\Omega_i \cap \Omega_j$ for any i, j, we have that there exists $0 = t_0 \le \cdots \le t_l \le \cdots \le t_K \le t_{K+1} = 1, l = 1, \ldots, K$ and corresponding $i_1, \ldots, i_K$ such that

$$h(t_l) = A_{i_l}x^{(l)} + c_{i_l} = A_{i_{l-1}}x^{(l)} + c_{i_{l-1}} \tag{4.12}$$

for all $1 \le l \le K$, where $\mathbf{x}^{(l)} = g(t_l)$. We also let $\mathbf{x}^{(0)} = g(t_0) = \mathbf{x}_1$ and $\mathbf{x}^{(K+1)} = g(t_{K+1}) = \mathbf{x}_2$. These $\mathbf{x}_l$'s have the property that

$$\sum_{l=0}^{K} \|\mathbf{x}_{i_{l+1}} - \mathbf{x}_{i_l}\|_\beta = \sum_{l=0}^{K}(t_{l+1} - t_l)\|\mathbf{x}_2 - \mathbf{x}_1\|_\beta = \|\mathbf{x}_2 - \mathbf{x}_1\|_\beta. \tag{4.13}$$

Thus we can bound the term $\|h(1) - h(0)\|_\alpha$ by

$$\begin{aligned}
\|h(1) - h(0)\|_\alpha &\leq \sum_{l=0}^{K} \|h(t_{l+1}) - h(t_l)\|_\alpha \\
&\leq \sum_{l=0}^{K} \|A_{i_{l+1}}\mathbf{x}_{l+1} + c_{i_{l+1}} - A_{i_l}\mathbf{x}_l - c_{i_l}\|_\alpha \\
&\overset{(i)}{\leq} \sum_{l=0}^{K-1} \|A_{i_l}\mathbf{x}_{l+1} + c_{i_l} - A_{i_l}\mathbf{x}_l - c_{i_l}\|_\alpha \\
&= \sum_{l=0}^{K} \|A_{i_l}\mathbf{x}_{l+1} - A_{i_l}\mathbf{x}_l\|_\alpha \\
&\overset{(ii)}{\leq} \sum_{l=0}^{K} \|A_{i_l}\|_{\alpha,\beta}\|\mathbf{x}_{l+1} - \mathbf{x}_l\|_\beta \\
&\leq \max_i\{\|A_i\|_{\alpha,\beta}, i \leq d\} \sum_{l=0}^{K} \|\mathbf{x}_{l+1} - \mathbf{x}_l\|_\beta \\
&\overset{(iii)}{\leq} \max_i\{\|A_i\|_{\alpha,\beta}, i \leq d\}\|\mathbf{x}_2 - \mathbf{x}_1\|_\beta = L,
\end{aligned}$$

where (i) is due to inequality (4.12), (ii) is due to the definition of operator norm, and (iii) is using equality (4.13). This proves inequality (4.10). □

Bibliography

1. Jason Abrevaya and Jian Huang, *On the bootstrap of the maximum score estimator*, Econometrica **73** (2005), no. 4, 1175–1204.
2. Dragi Anevski and Ola Hössjer, *A general asymptotic scheme for inference under order restrictions*, The Annals of Statistics **34** (2006), no. 4, 1874–1930.
3. Pramita Bagchi, Moulinath Banerjee, and Stilian A Stoev, *Inference for monotone functions under short-and long-range dependence: Confidence intervals and new universal limits*, Journal of the American Statistical Association **111** (2016), no. 516, 1634–1647.
4. Pramita Bagchi and Subhra Sankar Dhar, *A study on the least squares estimator of multivariate isotonic regression function*, Scandinavian Journal of Statistics **47** (2020), no. 4, 1192–1221.
5. Moulinath Banerjee and John A. Wellner, *Likelihood ratio tests for monotone functions*, The Annals of Statistics **29** (2001), no. 6, 1699–1731.
6. RERE Barlow, *Statistical inference under order restrictions; the theory and application of isotonic regression*, Tech. report, 1972.
7. Lawrence Brown Andreas Buja Berk, Richard, Edward George, et al., *Working with misspecified regression models*, Journal of Quantitative Criminology **34** (2018), 633–655.
8. Michael Best and Nilotpal Chakarvarti, *Active set algorithms for isotonic regression; a unifying framework*, Mathematical Programming **47** (1990), 425–439.
9. F Boussama, *Ergodicity, mixing and estimation in garch models*, Unpublished Ph. D. Dissertation, University of Paris **7** (1998).
10. H.D. Brunk, *Conditional expectation given a σ-lattice and applications*, The Annals of Statistics **36** (1965), 1339–1350.
11. Sabyasachi Chatterjee, Adityanand Guntuboyina, Bodhisattva Sen, et al., *On matrix estimation under monotonicity constraints*, Bernoulli **24** (2018), no. 2, 1072–1100.
12. Konstantinos Chatzikokolakis and Keye Martin, *A monotonicity principle for information theory*, Electronic Notes in Theoretical Computer Science **218** (2008), 111–129.
13. Yurii Aleksandrovich Davydov, *Mixing conditions for markov chains*, Teoriya Veroyatnostei i ee Primeneniya **18** (1973), no. 2, 321–338.
14. Subhra Sankar Dhar, *Trimmed mean isotonic regression*, Scandinavian Journal of Statistics **43** (2016), 202–212.

15. G. Ellison and S. Ellison, *Strategic entry deterrence and the behavior of pharmaceutical incumbents prior to patent expiration*, American Economic Journal: Microeconomics (2011), 1–36.
16. Arnaud Guyader, Nick Hengartner, Nicolas Jegou, and Eric Matzner-Lober, *Iterative isotonic regression*, ESAIM: Probability and Statistics **19** (2015), 1–23.
17. Qiyang Han, Tengyao Wang, Sabyasachi Chatterjee, and Richard J Samworth, *Isotonic regression in general dimensions*, The Annals of Statistics **47** (2019), no. 5, 2440–2471.
18. Linda Hoffmann, *Multivariate isotonic regression and its algorithms*, Ph.D. thesis, Wichita State University, 2009.
19. Ming-Hong Liu and Vasant Ubhaya, *Integer isotone optimization*, SIAM Journal on Optimization **7** (1997), 1152–1159.
20. Gary G Makowski, *Consistency of an estimator of doubly nondecreasing regression functions*, Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete **39** (1977), no. 4, 263–268.
21. R. Matzkin, *Restrictions of economic theory in nonparametric methods*, Handbook of Econometrics (R. Engle and D. McFadden, eds.), vol. IV, 1994, pp. 2523–2558.
22. Don L McLeish, *A maximal inequality and dependent strong laws*, The Annals of probability (1975), 829–839.
23. P. Milgrom and C. Shannon, *Monotone comparative statics*, Econometrica **62** (1994), 157–180.
24. Abdelkader Mokkadem, *Mixing properties of arma processes*, Stochastic processes and their applications **29** (1988), no. 2, 309–315.
25. Eric T Moolchan, Darrell L Hudson, Jennifer R Schroeder, and Shelley S Sehnert, *Heart rate and blood pressure responses to tobacco smoking among african-american adolescents.*, Journal of the National Medical Association **96** (2004), no. 6, 767.
26. W. F. Ramirez, *Computational methods for process simulation*, Elsevier Science, Oxford, 1989.
27. Emmanuel Rio, *Asymptotic theory of weakly dependent random processes*, vol. 80, Springer, 2017.
28. Tim Robertson and FT Wright, *Multiple isotonic median regression*, The Annals of Statistics (1973), 422–432.
29. W. Rudin, *Functional analysis*, International Series in Pure and Applied Mathematics, McGraw-Hill,, 1991.
30. Ou Zhao and Michael Woodroofe, *Estimating a monotone trend*, Statistica Sinica **22** (2012), no. 1, 359–378.

DEPARTMENT OF STATISTICS, VOLGENAU SCHOOL OF ENGINEERING, GEORGE MASON UNIVERSITY, USA.
Email address: pbagchi@gmu.edu

DEPARTMENT OF MATHEMATICS AND STATISTICS, IIT KANPUR, KANPUR, INDIA
Email address: subhra@iitk.ac.in