Electronic Journal of Statistics
Vol. 14 (2020) 2053–2097
ISSN: 1935-7524
https://doi.org/10.1214/20-EJS1714

The limiting behavior of isotonic and convex regression estimators when the model is misspecified

Eunji Lim

Decision Sciences
Adelphi University
1 South Ave, Garden City, NY 11530, USA
e-mail: elim@adelphi.edu

Abstract: We study the asymptotic behavior of the least squares estimators when the model is possibly misspecified. We consider the setting where we wish to estimate an unknown function $f_* : (0,1)^d \to \mathbb{R}$ from observations $(X,Y),(X_1,Y_1),\cdots,(X_n,Y_n)$; our estimator $\hat{g}_n$ is the minimizer of $\sum_{i=1}^n (Y_i - g(X_i))^2/n$ over $g \in \mathcal{G}$ for some set of functions $\mathcal{G}$. We provide sufficient conditions on the metric entropy of $\mathcal{G}$, under which $\hat{g}_n$ converges to g_* as $n \to \infty$, where g_* is the minimizer of $\|g - f_*\| \triangleq \mathbb{E}(g(X) - f_*(X))^2$ over $g \in \mathcal{G}$. As corollaries of our theorem, we establish $\|\hat{g}_n - g_*\| \to 0$ as $n \to \infty$ when $\mathcal{G}$ is the set of monotone functions or the set of convex functions. We also make a connection between the convergence rate of $\|\hat{g}_n - g_*\|$ and the metric entropy of $\mathcal{G}$. As special cases of our finding, we compute the convergence rate of $\|\hat{g}_n - g_*\|^2$ when $\mathcal{G}$ is the set of bounded monotone functions or the set of bounded convex functions.

MSC 2010 subject classifications: Primary 62G08, 62G20; secondary 62G10.

Keywords and phrases: Isotonic regression, convex regression, model misspecification, consistency, convergence rates, tests for misspecification.

Received July 2019.

Contents

1. Introduction

Given independent and identically distributed (iid) observations $(X, Y), (X_1, Y_1), \cdots, (X_n, Y_n)$, we consider the least squares estimator $\hat{g}_n$, which is the solution to the following problem:

$$\text{minimize } \varphi_n(f) \triangleq \frac{1}{n}\sum_{i=1}^{n}(Y_i - f(X_i))^2 \qquad \text{subject to } f \in \mathcal{G} \tag{1.1}$$

for a set $\mathcal{G}$ of functions, where

$$Y_i = f_*(X_i) + \varepsilon_i, \quad 1 \le i \le n,$$

for some unknown regression function $f_* : (0,1)^d \to \mathbb{R}$ satisfying $\mathbb{E}f_*(X)^2 < \infty$, and $((X_i, \varepsilon_i) : 1 \le i \le n)$ is a sequence of $(0,1)^d \times \mathbb{R}$–valued random vectors satisfying $\mathbb{E}(\varepsilon_1|X_1) = 0$ and $\mathbb{E}(\varepsilon_1^2|X_1) = \sigma^2 < \infty$. It should be emphasized that f_* does not need to belong to $\mathcal{G}$, so it is possible that the model is misspecified.

When $\mathcal{G}$ is the set of monotone functions, i.e.,

$$\mathcal{G} = \mathcal{G}_m \triangleq \{g : (0,1)^d \to \mathbb{R} : \mathbb{E}g(X)^2 < \infty \text{ and } g(x_1, \cdots, x_d) \le g(y_1, \cdots, y_d) \text{ if } x_j \le y_j \text{ for } 1 \le j \le d\},$$

the estimator $\hat{g}_n$ is known as the isotonic regression estimator. When $\mathcal{G}$ is the set of convex functions, i.e.,

$$\mathcal{G} = \mathcal{G}_c \triangleq \{g : (0,1)^d \to \mathbb{R} : \mathbb{E}g(X)^2 < \infty \text{ and } g \text{ is convex}\},$$

the estimator $\hat{g}_n$ is known as the convex regression estimator.

When studying the behavior of $\hat{g}_n$ as $n \to \infty$, the first question one needs to answer is, "What is the limit point of $(\hat{g}_n : n \ge 1)$?" To answer this question, we consider the space

$$L^2 = \{f : (0,1)^d \to \mathbb{R} : \mathbb{E}f(X)^2 < \infty\}.$$

We restrict our attention to the functions f satisfying $\mathbb{E}f(X)^2 < \infty$ since we expect $\varphi_n(f)$ to converge to $\mathbb{E}(Y - f(X))^2$ as $n \to \infty$. One such condition guaranteeing this is $\mathbb{E}f(X)^2 < \infty$ together with $\mathbb{E}(Y^2) < \infty$; see pages 198–199 of [53] for details. L^2 turns out to be a semi–Hilbert space equipped with the semi-inner product

$$(f_1, f_2) = \mathbb{E}(f_1(X)f_2(X))$$

and the associated seminorm

$$\|f\| = \sqrt{(f, f)}$$

for $f, f_1, f_2 \in L^2$. A semi-Hilbert space is a generalization of a Hilbert space where the inner product is required only to be a semi-inner product. A semi-inner product $\langle \cdot, \cdot \rangle$ is a generalization of an inner product where $\langle f, f \rangle = 0$ does

not necessarily imply $f = 0$. The fact that L^2 is a semi-Hilbert space plays an important role in the following discussion since it ensures the existence of a solution to (1.2).

Empirical L^2 seminorms $\|\cdot\|_n$ and $\|\cdot\|_{n,\delta}$ can be defined by

$$\|f\|_n = \left\{\frac{1}{n}\sum_{i=1}^n f(X_i)^2\right\}^{1/2}, \quad \|f\|_{n,\delta} = \left\{\frac{1}{n}\sum_{i=1}^n f(X_i)^2 I(X_i \in [\delta, 1-\delta]^d)\right\}^{1/2}$$

for $f \in L^2$ and $\delta > 0$.

A natural candidate for the limit point of $(\hat{g}_n : n \geq 1)$ in L^2 is the projection g_* of f_* onto $\mathcal{G}$, which is the solution to the following problem:

$$\text{minimize } \|g - f_*\|^2 \qquad \text{subject to } g \in \mathcal{G}. \tag{1.2}$$

Since L^2 is a semi-Hilbert space, under the assumption that $\mathcal{G}$ is a closed convex subset of L^2, a solution to (1.2) exists by the projection theorem and we will denote the set of solutions to (1.2) by $\mathcal{G}_*$. Of course, when $f_* \in \mathcal{G}$, we have $f_* \in \mathcal{G}_*$. In Theorem 3.1 of this paper, we will provide a set of sufficient conditions under which $\|\hat{g}_n - g_*\| \to 0$ as $n \to \infty$, thereby justifying $g_* \in \mathcal{G}_*$ as a limit point of $(\hat{g}_n : n \geq 1)$ in L^2.

When the model is well–specified, i.e., $f_* \in \mathcal{G}$, it is well-known that the metric entropy of $\mathcal{G}$ gives certain information about the consistency or convergence rate. (For definitions, see Section 2.) For example, in her unique work, [70] proved $\|\hat{g}_n - g_*\| \to 0$ under the conditions $f_* \in \mathcal{G}$, $\mathbb{E}\sup\{g(X)^2 : g \in \mathcal{G}\} < \infty$, and the ϵ-metric entropy of $\mathcal{G}$ in the $\|\cdot\|_n$ norm being $o_p(n)$. Furthermore, the metric entropy of $\mathcal{G}$ in the $\|\cdot\|$ or $\|\cdot\|_n$ norm provides explicit convergence rates for $\|\hat{g}_n - g_*\|$; see, for example, Theorem 4.1 of [71], Theorem 1 of [10], and Theorem 3.4.1 of [74].

However, when the model is misspecified, little is known about the relationship between the metric entropy of $\mathcal{G}$ and the behavior of $(\hat{g}_n : n \geq 1)$. The main objective of this paper is to establish a framework under which the metric entropy explains how $(\hat{g}_n : n \geq 1)$ behaves. The main contributions of this paper can be summarized as follows.

1. Our main theorem (Theorem 3.1) states that if the ϵ–metric entropy (without bracketing) of $\mathcal{G}$ in the $\|\cdot\|$ and $\|\cdot\|_{n,\delta}$ norm is finite almost surely (a.s.) for any $\epsilon > 0$, then
$$\|\hat{g}_n - g_*\| \to 0 \tag{1.3}$$
as $n \to \infty$ under the assumption that $\mathcal{G}$ is a closed convex subset of L^2. As corollaries of our theorem, we establish (1.3) when $\mathcal{G} = \mathcal{G}_m$ or $\mathcal{G} = \mathcal{G}_c$. So far, the results of type (1.3) have not been established when the model is misspecified; thus, (1.3) is one of the new contributions of this paper when $f_* \notin \mathcal{G}$. When the model is well–specified, the results of type (1.3) have been established by [70] under conditions such as the uniform boundedness of $g \in \mathcal{G}$ and the growth of the metric entropy of $\mathcal{G}$ in the $\|\cdot\|_n$ norm being of order $o_p(n)$. Our result can be viewed as an extension of some of the results in [70].

2. Our next finding on the rates of convergence (Theorem 4.1) states that when $\mathcal{G}$ is a closed convex subset of L^2, the ϵ-metric entropy with bracketing provides explicit convergence rates on $\|\hat{g}_n - g_*\|^2$. As corollaries of our finding, we obtain $\|\hat{g}_n - g_*\|^2 = \mathcal{O}_p(a_n)$, where

$$a_n = \begin{cases} n^{-2/3}, & \text{if } d = 1 \\ n^{-1/2}(\log n)^2, & \text{if } d = 2 \\ n^{-\frac{1}{2(d-1)}}, & \text{if } d > 2, \end{cases} \tag{1.4}$$

when

$$\mathcal{G} = \mathcal{G}_{m,B} \triangleq \{g \in \mathcal{G}_m : |g(x)| \leq B \text{ for } x \in (0,1)^d\}$$

for some constant B.
We also obtain $\|\hat{g}_n - g_*\|^2 = \mathcal{O}_p(b_n)$, where

$$b_n = \begin{cases} n^{-\frac{4}{4+d}}, & \text{if } d < 4 \\ (\log n)n^{-1/2}, & \text{if } d = 4 \\ n^{-2/d}, & \text{if } d > 4, \end{cases} \tag{1.5}$$

when

$$\mathcal{G} = \mathcal{G}_{c,B} \triangleq \{g \in \mathcal{G}_c : |g(x)| \leq B \text{ for } x \in (0,1)^d\}$$

for some constant B. We are not aware of any existing results in the literature on the rates of convergence when $d \geq 2$ and $f_* \notin \mathcal{G}$. Thus, the above results when $d \geq 2$ will shed light on the study of $(\hat{g}_n : n \geq 1)$ in the presence of model misspecification.

3. One of the implications of identifying the limit point of $(\hat{g}_n : n \geq 1)$ is that we are able to analyze the Type I error of a hypothesis test whose null and alternative hypotheses are given by

$$\begin{aligned} H_0 &: f_* \notin \mathcal{G} \\ H_a &: f_* \in \mathcal{G}. \end{aligned} \tag{1.6}$$

We will consider one of the popular test procedures, in which we observe $((X_i, Y_{ij}) : 1 \leq i \leq n, 1 \leq j \leq m)$, where $Y_{ij} = f_*(X_i) + \varepsilon_{ij}$. We then compute the test statistic as

$$\frac{m}{n}\sum_{i=1}^{n}(\tilde{Y}_i - \tilde{g}_n(X_i))^2, \tag{1.7}$$

where $\tilde{Y}_i = \sum_{j=1}^{m} Y_{ij}/m$, and $\tilde{g}_n$ is the solution to (1.1) with the Y_i's replaced by the $\tilde{Y}_i$'s. The finite sample behavior of this test statistic is well-known under the normality assumption on the ε_i's. The finite sample distribution of the test statistic is known as the chi-bar squared distribution; see, for example, (3.3) on page 51 of [68] and Section 2.1 of [66]. However, not much is known about the behavior of the test statistic when n and m are large. In this paper, we consider a very simple test procedure, which is based on the test statistic in (1.7), and we prove that its Type I error converges to 0 as $n \to \infty$ for m sufficiently large.

Shape constrained estimators have received considerable interest in the context of isotonic and convex regression estimators;

1. when $f_* \in \mathcal{G}$ and $d = 1$, some theoretical works on consistency and convergence rates can be found in [12], [6], [75], [59], [21], [71, 72], [58], [77], [26, 27], [14], [16], and [8] for isotonic regression, and [46], [44], [59], [35], [37], [9], [16], [15], and [8] for convex regression,
2. when$f_* \in \mathcal{G}$ and $d \geq 2$, consistency and convergence rates are studied by [43], [55], [17], [62], [10], and [41] for isotonic regression, and by [65] and [40] for convex regression,
3. when $f_* \notin \mathcal{G}$ and $d = 1$, some interesting results can be found in [16] and [8] for isotonic regression and in [37] and [8] for convex regression, and
4. when $f_* \notin \mathcal{G}$ and $d \geq 2$, we could not find any theoretical results concerning consistency or convergence rates. To the best of our knowledge, this paper is the first to establish consistency and compute the convergence rates when $f_* \notin \mathcal{G}$ and $d \geq 2$.

It should be noticed that our results are consistent with existing results when $f_* \in \mathcal{G}$ or $d = 1$. For example, let us consider the case where $\mathcal{G} = \mathcal{G}_{m,B}$. When $f_* \in \mathcal{G}_{m,B}$ and $d = 1$, the convergence rate in (1.4) agrees with the previous result given by [58] and [77]. When $f_* \notin \mathcal{G}_{m,B}$ and $d = 1$, the convergence rate in (1.4) is a slight improvement of the previous result established by [16], who computed the rate of $n^{-2/3}$ up to a logarithmic multiplicative factor in n; see Theorem 6.1 of [16]. When $f_* \in \mathcal{G}_{m,B}$ and $d = 2$, the convergence rate in (1.4) is a slight improvement of the previous result established by [17], who computed the rate of $n^{-1/2}(\log n)^4$; see Theorem 2.1 of [17]. When $f_* \in \mathcal{G}_{m,B}$ and $d \geq 3$, the convergence rate in (1.4) is slower than the previous result established by [41], who computed the rate of $n^{-1/d}(\log n)^4$; see Theorem 1 of [41]. Next, let us consider the case where $\mathcal{G} = \mathcal{G}_{c,B}$. When $f_* \in \mathcal{G}_{c,B}$ and $d = 1$, the convergence rate in (1.5) agrees with the previous results obtained by [35], [15], and [40]. When $f_* \notin \mathcal{G}_{c,B}$ and $d = 1$, the convergence rate in (1.5) is a slight improvement of the result obtained by [37], who achieved the convergence rate of $n^{-4/5}(\log n)^{5/4}$. When $f_* \in \mathcal{G}_{c,B}$ and $d \geq 2$, the convergence rate in (1.5) agrees with the previous result obtained by [40]; see Theorem 3.6 of [40].

Related works can be found in [6], [28], [63], [29], [50], [52], [42], [51], and [69] for computational algorithms used to compute isotonic or convex regression estimators; in [34], [61], [39], [60], [22], [2], [67], [49], [64], [19], [20], [13], [24], [3], [23], [25], and [48] for the density estimation under shape restriction; in [1], [45], [56], [57], and [18] for additive models; and in [7], [68], [76], [38], [32], [4], [66], and [5] for hypothesis tests to detect monotonicity and convexity.

This paper is organized as follows. In Section 2, we introduce some notations and definitions. Section 3 presents Theorem 3.1, our main theorem on consistency, and proves (1.3) for the cases of $\mathcal{G} = \mathcal{G}_m$ and $\mathcal{G} = \mathcal{G}_c$ as corollaries. In Section 4, we describe Theorem 4.1, our main finding on the convergence rate, and obtain (1.4) and (1.5) for the cases of $\mathcal{G} = \mathcal{G}_{m,B}$ and $\mathcal{G} = \mathcal{G}_{c,B}$ as corollaries. Section 5 considers a test procedure for testing $f_* \in \mathcal{G}$, analyzes its Type 1 error, and examine its numerical behavior.

2. Notation and Definitions

For $x \in \mathbb{R}^d$, we write its jth component as x_j. Thus, $x = (x_1, \cdots, x_d)$ and its norm is given by $|x| = \{\sum_{j=1}^n x_j^2\}^{1/2}$. The transpose of x is denoted by x^T. For $a, b \in \mathbb{R}^d$, we write $a \leq b$ if $a_i \leq b_j$ for $1 \leq j \leq d$. For $a, b \in \mathbb{R}^d$, the hyperrectangle $[a, b]$ is the set $\{x \in \mathbb{R}^d : a \leq x \leq b\}$. For $x \in \mathbb{R}$, $\lceil x \rceil$ is the smallest integer that is greater than or equal to x. $[x]^+$ equals x if $x \geq 0$, and 0 otherwise. $[x]^-$ equals $-x$ if $x \leq 0$, and 0 otherwise. For $x, y \in \mathbb{R}$, $x \wedge y = x$ if $x \leq y$ and y otherwise. For sequences of real numbers $(\alpha_n : n \geq 1)$ and $(\beta_n : n \geq 1)$, we write $\alpha_n \lesssim \beta_n$ if $\alpha_n \leq C\beta_n$ for some constant C and all $n \geq 1$. For a sequence of random variables $(Z_n : n \geq 1)$ and a sequence of positive real numbers $(\alpha_n : n \geq 1)$, we say $Z_n = \mathcal{O}_p(\alpha_n)$ as $n \to \infty$ if, for any $\epsilon > 0$, there exist constants C and N such that $\mathbb{P}(|Z_n/\alpha_n| > C) < \epsilon$ for $n \geq N$. We also say $Z_n = o_p(\alpha_n)$ as $n \to \infty$ if Z_n/α_n converges to zero in probability as $n \to \infty$.

Covering and Bracketing Numbers

Let d be a pseudo-metric on a set $\mathcal{A}$ of functions. A set $\mathcal{B} \subset \mathcal{A}$ is called an ϵ-net for $(\mathcal{A}, d)$ if for each $f \in \mathcal{A}$, there exists $h \in \mathcal{B}$ such that $d(f, h) \leq \epsilon$. For $\epsilon > 0$, the covering number $N(\epsilon, \mathcal{A}, d)$ is defined as the minimal number of elements in an ϵ–net for $(\mathcal{A}, d)$. In other words,

$$N(\epsilon, \mathcal{A}, d) \triangleq \inf\{J : \text{ There exist } h_1, \cdots, h_J \in \mathcal{A} \text{ such that } \{h_1, \cdots, h_J\} \text{ is an } \epsilon\text{–net for } (\mathcal{A}, d)\}.$$

We set $N(\epsilon, \mathcal{A}, d) = \infty$ if no ϵ–net exists. The ϵ–metric entropy for $(\mathcal{A}, d)$ is $\log(1 + N(\epsilon, \mathcal{A}, d))$.

Given $l, u \in \mathcal{A}$, the bracket $[l, u]$ is the set of functions f satisfying $l \leq f \leq u$. An ϵ-bracket is a bracket $[l, u]$ satisfying $d(l, u) \leq \epsilon$. For $\epsilon > 0$, the bracketing number $N_{[\,]}(\epsilon, \mathcal{A}, d)$ is the minimum number of ϵ–brackets needed to cover $\mathcal{A}$. In other words,

$$N_{[\,]}(\epsilon, \mathcal{A}, d) \triangleq \inf\{J : \text{ There exist } l_1, \cdots, l_J, u_1, \cdots, u_J \in \mathcal{A} \text{ such that } d(l_j, u_j) \leq \epsilon \text{ for } 1 \leq j \leq J \text{ and } \mathcal{A} \subset \bigcup_{j=1}^J [l_j, u_j]\}.$$

The ϵ–metric entropy with bracketing for $(\mathcal{A}, d)$ is $\log(1 + N_{[\,]}(\epsilon, \mathcal{A}, d))$.

Metrics

Let $\mathcal{A}$ be a set of functions defined on $(0, 1)^d$. We define a metric d_2 on $\mathcal{A}$ as follows:

$$d_1(f_1, f_2) = \int_{(0,1)^d} |f_1(x) - f_2(x)| dx$$

$$d_2(f_1, f_2) = \left\{ \int_{(0,1)^d} (f_1(x) - f_2(x))^2 dx \right\}^{1/2}$$

For $\delta > 0$, let $A(\delta)$ be a set of functions defined on $[\delta, 1-\delta]^d$. We define a pseudo-metric d_n^δ and metrics d_2^δ and d_∞^δ on $\mathcal{A}(\delta)$ as follows:

$$d_n^\delta(f_1, f_2) = \left\{ \frac{1}{n} \sum_{i=1}^n (f_1(X_i) - f_2(X_i))^2 I(X_i \in [\delta, 1-\delta]^d) \right\}^{1/2}$$

$$d_2^\delta(f_1, f_2) = \left\{ \int_{[\delta,1-\delta]^d} (f_1(x) - f_2(x))^2 dx \right\}^{1/2}$$

$$d_\infty^\delta(f_1, f_2) = \sup_{x \in [\delta,1-\delta]^d} |f_1(x) - f_2(x)|.$$

It should be noted that $N(\epsilon, \mathcal{A}(\delta), d_n^\delta)$ depends on $X_1, \cdots, X_n$, so it is a random variable.

3. Main Theorem on Consistency

In this section, we precisely describe our main theorem on consistency, Theorem 3.1. Let us consider the following problem:

$$\text{minimize } \varphi_n(f) \triangleq \frac{1}{n} \sum_{i=1}^n (Y_i - f(X_i))^2 \qquad \text{subject to } f \in \mathcal{G}. \tag{3.1}$$

In general, there may not exist a solution to (3.1). However, in some specific cases such as the case of $\mathcal{G} = \mathcal{G}_m$ or $\mathcal{G} = \mathcal{G}_c$, a solution to (3.1) exists. We will first develop a general theory under the assumption that there exists a solution to (3.1). We will then consider the cases of $\mathcal{G} = \mathcal{G}_m$ and $\mathcal{G} = \mathcal{G}_c$, and establish corollaries.

In order to analyze the behavior of the solution to (3.1), we need the following assumptions:

A1. $(X, Y), (X_1, Y_1), (X_2, Y_2), \cdots$ is a sequence of iid $(0,1)^d \times \mathbb{R}$–valued random vectors satisfying

$$Y = f_*(X) + \varepsilon \text{ and } Y_i = f_*(X_i) + \varepsilon_i \text{ for } i \geq 1.$$

A2. $X, X_1, X_2, \cdots$ have a common positive density function $\tau : (0,1)^d \to \mathbb{R}$.
A3. There exists a positive constant τ_* such that $\tau(x) \leq \tau_*$ for $x \in (0,1)^d$.
A4. (i) $\mathbb{E}(\varepsilon|X) = \mathbb{E}(\varepsilon_i|X_i) = 0$ and (ii) $\mathbb{E}(\varepsilon^2|X) = \mathbb{E}(\varepsilon_i^2|X_i^2) = \sigma^2 < \infty$ for $i \geq 1$.
A5. $\mathbb{E} f_*(X)^2 < \infty$.

Under A1–A5, several properties are satisfied when $\mathcal{G} = \mathcal{G}_m$ or $\mathcal{G} = \mathcal{G}_c$. One of such properties is presented next.

P1. $\mathcal{G}$ is a closed convex subset of L^2.

With A5 and P1 in force, the projection theorem guarantees the existence of a solution to the following problem:

$$\text{minimize } \|f - f_*\| \qquad \text{subject to } f \in \mathcal{G}. \tag{3.2}$$

A solution g_* to (3.2) exists and is unique up to a set of measure zero under A2. Furthermore, g_* is characterized by the following relationship:

$$(f_* - g_*, g - g_*) \leq 0 \tag{3.3}$$

for $g \in \mathcal{G}$; see, for example, Theorem 1 on page 69 of [54]. We will denote the set of the solutions to (3.2) by $\mathcal{G}_*$.

We next present other properties that are satisfied when $\mathcal{G} = \mathcal{G}_m$ or $\mathcal{G} = \mathcal{G}_c$. In P3, we focus on a subset of $\mathcal{G}$, to which the solution to (3.1) belongs for n sufficiently large. P3 states that when one restricts the functions in such a subset on $[\delta, 1-\delta]$ for $\delta > 0$, there exist ϵ-nets for such restricted functions in the d_2^δ and d_n^δ pseudo-metrics, respectively, and the ϵ-net in the d_n^δ metric is independent of n for n sufficiently large. P4 states that for any $\delta > 0$, g_* and the solution to (3.1) are uniformly Lipschitz over $[\delta, 1-\delta]$ and for n sufficiently large.

P2. For $n \geq 1$, there exists a solution $\hat{g}_n$ to (3.1). We will denote the set of the solutions to (3.1) by $\hat{\mathcal{G}}_n$.

P3. For each $\delta > 0$, there exists a subset $\mathcal{H}(\delta) \subset \mathcal{G}$ such that

$$\mathbb{P}(\hat{\mathcal{G}}_n \subset \mathcal{H}(\delta) \text{ for all but finitely many } n) = 1. \tag{3.4}$$

Let $\tilde{\mathcal{H}}(\delta)$ be the set of functions in $\mathcal{H}(\delta)$ restricted to $[\delta, 1-\delta]^d$, i.e.,

$$\tilde{\mathcal{H}}(\delta) \triangleq \{\tilde{h} : [\delta, 1-\delta]^d \to \mathbb{R} : \text{There exists } h \in \mathcal{H}(\delta) \text{ such that } \tilde{h}(x) = h(x) \text{ for } x \in [\delta, 1-\delta]^d\}.$$

(i) For any $\epsilon > 0$ and $\delta > 0$, $N(\epsilon, \tilde{\mathcal{H}}(\delta), d_2^\delta) < \infty$.
(ii) For any $\epsilon > 0$ and $\delta > 0$, there exist a constant $r \triangleq r(\epsilon, \delta)$ and $\{h_1, \cdots, h_r\} \subset \tilde{\mathcal{H}}(\delta)$ such that

$$\mathbb{P}(\limsup_{n\to\infty} \sup_{g \in \tilde{\mathcal{H}}(\delta)} \min_{1 \leq j \leq r} d_n^\delta(g, h_j) \leq \epsilon) = 1. \tag{3.5}$$

P4. For any $\delta > 0$, there exists a constant $\beta(\delta)$ such that

$$\mathbb{P}\Bigg(\sup_{g_* \in \mathcal{G}_*, x, y \in [\delta, 1-\delta]^d, x \neq y} |g_*(x) - g_*(y)|/|x - y| \leq \beta(\delta) \text{ and } \sup_{\hat{g}_n \in \hat{\mathcal{G}}_n, x, y \in [\delta, 1-\delta]^d, x \neq y} |\hat{g}_n(x) - \hat{g}_n(y)|/|x - y| \leq \beta(\delta) \text{ for all but finitely many } n \Bigg) = 1.$$

Remark 3.1. *The requirement described in P3(ii) is satisfied if, for any δ, we have $N(\epsilon, \tilde{\mathcal{H}}(\delta), d_n^\delta) < \infty$ a.s. for n sufficiently large and there exists an ϵ-net for $(\tilde{\mathcal{H}}(\delta), d_n^\delta)$ that does not depend on n for n sufficiently large.*

Remark 3.2. *The requirement described in P3(ii) is satisfied if, for any δ, we have $N(\epsilon, \tilde{\mathcal{H}}(\delta), d_\infty^\delta) < \infty$ a.s. for n sufficiently large.*

We are ready to present our main result regarding consistency, whose proof is provided in Appendix A.

Theorem 3.1. *Assume A1–A5 and P1–P3. Then, for any $g_* \in \mathcal{G}_*$ and $\hat{g}_n \in \hat{\mathcal{G}}_n$,*

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{g}_n(X_i) - g_*(X_i))^2 \to 0 \text{ a.s. as } n \to \infty \text{ and} \tag{3.6}$$

$$\|\hat{g}_n - g_*\| \to 0 \text{ as } n \to \infty. \tag{3.7}$$

If, in addition, P4 holds, then for any $\delta > 0$, $g_* \in \mathcal{G}_*$, and $\hat{g}_n \in \hat{\mathcal{G}}_n$

$$\sup_{x \in [\delta, 1-\delta]^d} |\hat{g}_n(x) - g_*(x)| \to 0 \tag{3.8}$$

a.s. as $n \to \infty$.

Corollary 3.1. *Let $\mathcal{G} = \mathcal{G}_m \triangleq \{g \in L^2 : g(x_1, \cdots, x_d) \le g(y_1, \cdots, y_d)$ if $x_j \le y_j$ for $1 \le j \le d\}$. Under A1–A5, P1–P3 are satisfied, and (3.6) and (3.7) hold. If, in addition, $d = 1$ and f_* is continuous, then the solution to (3.2) exists uniquely and (3.8) holds.*

Corollary 3.2. *Let $\mathcal{G} = \mathcal{G}_c \triangleq \{g \in L^2 : g$ is convex.$\}$. Under A1–A5, P1–P4 are satisfied, the solution to (3.2) exists uniquely, and (3.6), (3.7), and (3.8) hold.*

Remark 3.3. *Results similar to (3.7) have been established by [8], [16], and [37]. Also, an analysis of type (3.8) exists when the model is well-specified; see, for example, [62] for the case of isotonic regression and [65] and [53] for the case of convex regression. However, when the model is misspecified, the only work in the form of (3.8) is done by [53], who established (3.8) when $\mathcal{G}$ is the set of convex functions bounded by some known function $k_* \in L^2$. In Corollary 3.2, we are able to drop the requirement that $|g| \le k_*$ for $g \in \mathcal{G}$.*

Corollary 3.3. *Let $\mathcal{G} = \mathcal{G}_{m,B} \triangleq \{g \in \mathcal{G}_m : |g(x)| \le B$ for $x \in (0,1)^d\}$ for some constant $B > 0$. Under A1–A5, P1–P3 are satisfied, and (3.6) and (3.7) hold. If, in addition, $d = 1$ and f_* is continuous, then the solution to (3.2) exists uniquely and (3.8) holds.*

Corollary 3.4. *Let $\mathcal{G} = \mathcal{G}_{c,B} \triangleq \{g \in \mathcal{G}_c : |g(x)| \le B$ for $x, y \in (0,1)^d\}$ for some constant $B > 0$. Under A1–A5, P1–P4 are satisfied, the solution to (3.2) exists uniquely, and (3.6), (3.7), and (3.8) hold.*

The proofs of Corollaries 3.1–3.4 are given in Appendix A.

In P3, we focus on the set of functions $g \in \mathcal{G}$ restricted on $[\delta, 1-\delta]^d$, and impose some conditions on the ϵ-metric entropy of such a set. We consider the functions restricted on $[\delta, 1-\delta]^d$ because whether $\hat{g}_n \in \hat{\mathcal{G}}_n$ is stochastically bounded near the boundary of $(0,1)^d$ is not fully answered. For isotonic regression with Gaussian errors, this boundary issue is tackled in [41]. [41] establishes convergence rates in the setting of isotonic regression without assuming any uniform bound on functions $g \in \mathcal{G}$. However, whether or not the convex regression estimator without a uniform bound has stochastically bounded behavior near the boundary of $(0,1)^d$ is still an open question.

We close this section by introducing a set of conditions that can greatly simplify the problem. To establish consistency, we use the following inequality

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n}(\hat{g}_n(X_i)-g_*(X_i))^2 &\leq \frac{2}{n}\sum_{i=1}^{n}(Y_i-g_*(X_i))(\hat{g}_n(X_i)-g_*(X_i)) \\ &= \frac{2}{n}\sum_{i=1}^{n}(f_*(X_i)-g_*(X_i))(\hat{g}_n(X_i)-g_*(X_i)) \\ &+ \frac{2}{n}\sum_{i=1}^{n}\varepsilon_i(\hat{g}_n(X_i)-g_*(X_i)) \end{aligned} \tag{3.9}$$

as in [73]. In the presence of the following Glivenko–Cantelli type conditions, (3.6) and (3.7) can be easily established.

P5. (i)

$$\sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{i=1}^{n}(f_*(X_i)-g_*(X_i))g(X_i)-\mathbb{E}(f_*(X)-g_*(X))g(X)\right|\to 0$$

a.s. as $n\to\infty$.
(ii)

$$\sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i g(X_i)-\mathbb{E}(\varepsilon g(X))\right|\to 0 \text{ a.s. as } n\to\infty.$$

P6.

$$\sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{i=1}^{n}(g(X_i)-g_*(X_i))^2-\mathbb{E}(g(X)-g_*(X))^2\right|\to 0 \text{ a.s. as } n\to\infty.$$

The following theorem establishes (3.6) and (3.7) under P5 and P6.

Theorem 3.2. *Assume A1, A4, A5, P1, P2, and P5. Then, for any $g_* \in \mathcal{G}_*$ and $\hat{g}_n \in \hat{\mathcal{G}}_n$, (3.6) holds. If, in addition, P6 is satisfied, then (3.7) holds.*

Proof of Theorem 3.2. Combine (3.3), (3.9), and the strong law of large numbers. □

One of the important implications of Theorem 3.2 is that, in the presence of A1–A5 and P1–P2, P3 in Theorem 3.1 can be replaced by the following conditions on the bracketing number and the uniform boundedness of f_*, $g \in \mathcal{G}$, and ε:

P7.

(i) $N_{[\,]}(\epsilon, \mathcal{G}, d_1) < \infty$ for every $\epsilon > 0$.

(ii) There exists $M_* \geq 0$ such that

$$|f_*(x)| \leq M_* \text{ for all } x \in (0,1)^d$$
$$|g(x)| \leq M_* \text{ for all } x \in (0,1)^d \text{ and } g \in \mathcal{G}$$
$$|\varepsilon| \leq M_* \text{ a.s.}$$

To see why P7 can replace P5 and P6, let $\mathcal{F} = \{(f_* - g_*)g : g \in \mathcal{G})\}$. For any $g_1, g_2 \in \mathcal{G}$, we have

$$d_1((f_* - g_*)g_1, (f_* - g_*)g_2) = \int_{(0,1)^d} |f_*(x) - g_*(x)||g_1(x) - g_2(x)|dx \leq 2M_* d_1(g_1, g_2). \quad (3.10)$$

Hence, P7(i) and (3.10) imply $N_{[\,]}(\epsilon, \mathcal{F}, d_1) < \infty$ for every $\epsilon > 0$, and Theorem 2.4.1 on page 122 of [74] establishes P5(i). Similar arguments can be applied to establish P5(ii) and P6.

The above arguments prove the following theorem.

Theorem 3.3. *Assume A1–A5 and P1–P2. Furthermore, assume P7. Then, for any $g_* \in \mathcal{G}_*$ and $\hat{g}_n \in \hat{\mathcal{G}}_n$, (3.6) and (3.7) hold.*

The following corollaries establish consistency of $\hat{g}_n$ for the case when $\mathcal{G} = \mathcal{G}_{m,B}$ and $\mathcal{G} = \mathcal{G}_{c,B}$ under P7.

Corollary 3.5. *Let $\mathcal{G} = \mathcal{G}_{m,B}$ be defined as in Corollary 3.3. Assume A1–A5. Furthermore, assume that f_* is uniformly bounded and ε is bounded, i.e., there exist constants M_1 and M_2 such that $|f_*(x)| \leq M_1$ for all $x \in (0,1)^d$ and $|\varepsilon| \leq M_2$ a.s. Then, P1, P2, and P7 are satisfied, and hence, (3.6) and (3.7) hold for any $g_* \in \mathcal{G}_*$ and $\hat{g}_n \in \hat{\mathcal{G}}_n$.*

Proof of Corollary 3.5. P1 and P2 can be established using the arguments similar to those in Steps 1 and 2 in the Proof of Corollary 3.1. P7(i) is established by Theorem 1.1 of [30]. Theorem 3.3 then implies the desired conclusions. □

Corollary 3.6. *Let $\mathcal{G} = \mathcal{G}_{c,B}$ be defined as in Corollary 3.4. Assume A1–A5. Furthermore, assume that f_* is uniformly bounded and ε is bounded, i.e., there exist constants M_1 and M_2 such that $|f_*(x)| \leq M_1$ for all $x \in (0,1)^d$ and $|\varepsilon| \leq M_2$ a.s. Then, P1, P2, and P7 are satisfied, and hence, (3.6) and (3.7) hold for any $g_* \in \mathcal{G}_*$ and $\hat{g}_n \in \hat{\mathcal{G}}_n$.*

Proof of Corollary 3.6. P1 and P2 can be established using the arguments similar to those in Steps 1 and 2 in the Proof of Corollary 3.2. P7(i) is established by Theorem 1.1(ii) on page 567 of [31]. Theorem 3.3 then implies the desired conclusions. □

4. Convergence Rates

By identifying the limit point of $(\hat{g}_n : n \geq 1)$ as g_*, we are now able to obtain the rate at which $\mathbb{E}(\hat{g}_n(X) - g_*(X))^2$ converges to 0. We start by rewriting

$$\begin{aligned} Y &= f_*(X) + \varepsilon \\ &= g_*(X) + (f_*(X) - g_*(X) + \varepsilon) \end{aligned}$$

and treating $f_*(X) - g_*(X) + \varepsilon$ as an error term around g_*. The problem is now converted into the question of how to find an unknown function g_* (rather than f_*) when the observations come with errors in the form $f_*(X) - g_*(X) + \varepsilon$, whose mean is not zero. Our estimator $\hat{g}_n$ minimizes

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n}(Y_i - g(X_i))^2 &= \frac{1}{n}\sum_{i=1}^{n}((f_*(X_i) - g_*(X_i) + \varepsilon_i) + (g_*(X_i) - g(X_i)))^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(f_*(X_i) - g_*(X_i) + \varepsilon_i)^2 \\ &\quad -\frac{2}{n}\sum_{i=1}^{n}(f_*(X_i) - g_*(X_i) + \varepsilon_i)(g(X_i) - g_*(X_i)) + \frac{1}{n}\sum_{i=1}^{n}(g(X_i) - g_*(X_i))^2 \end{aligned}$$

over $g \in \mathcal{G}$, so it maximizes

$$\mathbb{M}_n(g) \triangleq \frac{2}{n}\sum_{i=1}^{n}(f_*(X_i) - g_*(X_i) + \varepsilon_i)(g(X_i) - g_*(X_i)) - \frac{1}{n}\sum_{i=1}^{n}(g(X_i) - g_*(X_i))^2$$

over $g \in \mathcal{G}$. Many existing theories on the convergence rates of least squares estimators do not require mean zero errors. Instead, they require $\mathbb{E}\mathbb{M}_n(g) - \mathbb{E}\mathbb{M}_n(g_*)$ to have a negative drift. In our case, this is satisfied because

$$\begin{aligned} &\mathbb{E}\mathbb{M}_n(g) - \mathbb{E}\mathbb{M}_n(g_*) \\ &\quad = 2\mathbb{E}(f_*(X) - g_*(X))(g(X) - g_*(X)) - \mathbb{E}(g(X) - g_*(X))^2 \\ &\quad \leq -\mathbb{E}(g(X) - g_*(X))^2 \end{aligned} \tag{4.1}$$

due to the fact that $\mathcal{G}$ is a closed convex subset of L^2, and hence, (3.3) is satisfied. (4.1) enables us to utilize existing theories on the convergence rates without having mean zero errors around g_*.

We start with some additional assumptions.

- A6. f_* is uniformly bounded, i.e., there exists a constant A_* such that $|f_*(x)| \leq A_*$ for $x \in (0,1)^d$.
- A7. The functions in $\mathcal{G}$ are uniformly bounded, i.e., there exists a constant B_* such that $|g(x)| \leq B_*$ for $x \in (0,1)^d$ and $g \in \mathcal{G}$.
- A8. The ε_i's satisfy $\mathbb{E}(\varepsilon_i | X_i) = 0$ for $i \geq 1$ and they are subexponential, i.e., there exist constants γ and Γ satisfying $\mathbb{E}\exp(\Gamma|\varepsilon|) \leq \gamma$.

The following theorem is our main tool for computing convergence rates and is a modification of Theorem 3.4.1 of [74].

Theorem 4.1. *Assume A1–A3, A6–A8, and P1–P2. For each $n \geq 1$, define $\phi_n : (0, \infty) \to [0, \infty]$ by*

$$\phi_n(t) = \tilde{J}_{[\,]}(t, \mathcal{G}, d_2)(1 + \tilde{J}_{[\,]}(t, \mathcal{G}, d_2)/(t^2\sqrt{n}))$$

for $t \geq 0$, where

$$\tilde{J}_{[\,]}(t, \mathcal{G}, d_2) \triangleq \int_{(t^2/64)\wedge(t/24)}^{t} \sqrt{1 + \log N_{[\,]}(u, \mathcal{G}, d_2)}du.$$

Suppose $\delta \mapsto \phi_n(\delta)/\delta^\alpha$ is decreasing on $(0, \infty)$, for some $\alpha < 2$. Let r_n satisfy $r_n^2\phi_n(1/r_n) \lesssim \sqrt{n}$ for every n. Then

$$\|\hat{g}_n - g_*\|^2 = \mathcal{O}_p(1/r_n^2)$$

as $n \to \infty$.

The proof of Theorem 4.1 is deferred to Appendix A. As corollaries of Theorem 4.1, we obtain the convergence rates when $\mathcal{G} = \mathcal{G}_{m,B}$ and $\mathcal{G} = \mathcal{G}_{c,B}$, respectively. The proofs of Corollaries 4.1 and 4.2 are provided in Appendix A.

Corollary 4.1. *Let $\mathcal{G} = \mathcal{G}_{m,B}$ be defined as in Corollary 3.3. Under A1–A7,*

$$\|\hat{g}_n - g_*\|^2 = \mathcal{O}_p(a_n) \quad \text{as } n \to \infty,$$

where

$$a_n = \begin{cases} n^{-2/3}, & \text{if } d = 1 \\ n^{-1/2}(\log n)^2, & \text{if } d = 2 \\ n^{-\frac{1}{2(d-1)}}, & \text{if } d > 2. \end{cases} \tag{4.2}$$

Remark 4.1 (Comparison with Existing Works When $d = 1$)**.** *When the model is well-specified, i.e., $f_* \in \mathcal{G}$, the convergence rate obtained in Corollary 4.1 when $d = 1$ agrees with the previous result given by [58] and [77], who established the rate of $n^{-2/3}$ for the univariate isotonic regression estimators; see Theorem 2.2 of [77]. On the other hand, when the model is misspecified, i.e., $f_* \notin \mathcal{G}$, the convergence rate obtained in Corollary 4.1 when $d = 1$ is a slight improvement of the previous result established by [16], who computed the rate of $n^{-2/3}$ up to a logarithmic multiplicative factor in n; see Theorem 6.1 of [16].*

Remark 4.2 (Comparison with Existing Works When $d = 2$)**.** *When the model is well–specified, i.e., $f_* \in \mathcal{G}$, then the convergence rate obtained in Corollary 4.1 when $d = 2$ is a slight improvement of the previous result established by [17], who computed the rate of $n^{-1/2}(\log n)^4$; see Theorem 2.1 of [17]. When the model is misspecified and $d = 2$, we could not find any existing theoretical work on the convergence rates.*

Remark 4.3 (Comparison with Existing Works When $d \geq 3$)**.** *When the model is well–specified, i.e., $f_* \in \mathcal{G}$, then the convergence rate obtained in Corollary 4.1 when $d \geq 3$ is slower than the previous result established by [41], who computed the rate of $n^{-1/d}(\log n)^4$; see Theorem 1 of [41]. Corollary 4.1 computes suboptimal rates compared with those obtained in [41] since the entropy integral, defined by $\int_0^t \sqrt{1 + \log N_{[\,]}(u, \mathcal{G}, d_2)}du$ for $t > 0$, is divergent when $d \geq 3$. When the model is misspecified and $d \geq 3$, we could not find any existing theoretical work on the convergence rates.*

Corollary 4.2. *Let $\mathcal{G} = \mathcal{G}_{c,B}$ be defined as in Corollary 3.4. Under A1–A7,*

$$\|\hat{g}_n - g_*\|^2 = \mathcal{O}_p(b_n) \quad \text{as } n \to \infty,$$

where

$$b_n = \begin{cases} n^{-\frac{4}{4+d}}, & \text{if } d < 4 \\ (\log n)n^{-1/2}, & \text{if } d = 4 \\ n^{-2/d}, & \text{if } d > 4. \end{cases} \tag{4.3}$$

Remark 4.4 (Comparison with Existing Works When $d = 1$)**.** *When the model is well-specified, the convergence rate obtained in Corollary 4.2 when $d = 1$ agrees with the previous results obtained by [35], [15], and [40]. On the other hand, when the model is misspecified, the rate in Corollary 4.2 when $d = 1$ is a slight improvement of the result obtained by [37], who achieved the convergence rate of $n^{-4/5}(\log n)^{5/4}$; see Theorem 6.1 of [37].*

Remark 4.5 (Comparison with Existing Works When $d \geq 2$)**.** *When the model is well–specified, the convergence rate obtained in Corollary 4.2 when $d \geq 2$ agrees with the previous result obtained by [40]; see Theorem 3.6 of [40]. When the model is misspecified and $d \geq 2$, we could not find any existing theoretical work on the convergence rates.*

5. A Test for Misspecification

The case of model misspecification is of particular interest to practitioners because even though $\hat{g}_n$ is a nice estimator of f_* when $f_* \in \mathcal{G}$, it is typically not known a priori whether the function f_* truly belongs to $\mathcal{G}$ or not. Thus, a modeler must conduct a hypothesis test in order to be confident about $f_* \in \mathcal{G}$ before he or she suggests $\hat{g}_n$ as an estimator of f_*. In this section, we introduce a test procedure for detecting whether f_* belongs to $\mathcal{G}$ or not. We consider the following null and alternative hypotheses:

$$H_0 : f_* \notin \mathcal{G}$$
$$H_a : f_* \in \mathcal{G}.$$

In our test procedure, we assume that f_* can be observed multiple times at any given design points $(X_i : 1 \leq i \leq n)$, and hence, the observed data can be expressed as $((X_i, Y_{ij}) : 1 \leq i \leq n, 1 \leq j \leq m)$, where

$$Y_{ij} = f_*(X_i) + \varepsilon_{ij}.$$

In Section 5.1, we precisely describe the test procedure under consideration. In Proposition 5.1, we establish that the Type I error of the test procedure converges to 0 as $n \to \infty$ for m sufficiently large. Section 5.2 is devoted to extensive simulation studies, which demonstrate that the test procedure successfully detects the non–monotonicity or non–convexity of test functions as $n \to \infty$.

5.1. Test Procedure under Consideration

The key idea of the test procedure is to observe that $\tilde{Y}_i = \sum_{j=1}^m Y_{ij}/m$ converges to $f_*(X_i)$ as $m \to \infty$ for each $i \in \{1, \cdots, n\}$ and fixed n. We define $\tilde{g}_n$ by the solution to (3.1) with the Y_i's replaced by the $\tilde{Y}_i$'s. Let n be fixed and $m \to \infty$. In such a situation, we can expect $\tilde{Y}_i$ to converge to $f_*(X_i)$ for $1 \le i \le n$. We also expect $\tilde{g}_n(X_i)$ to converge to $g_*(X_i)$ for $1 \le i \le n$. Therefore, if $f_* \in \mathcal{G}$, we expect $\sum_{i=1}^n (\tilde{Y}_i - \tilde{g}_n(X_i))^2/n$ to converge to 0, whereas if $f_* \notin \mathcal{G}$, we expect $\sum_{i=1}^n (\tilde{Y}_i - \tilde{g}_n(X_i))^2/n$ to converge to $\sum_{i=1}^n (f_*(X_i) - g_*(X_i))^2/n$, which is a positive number. Inspired by this observation, we consider the following test statistic:

$$\mathrm{TS} = \frac{m}{n} \sum_{i=1}^n (\tilde{Y}_i - \tilde{g}_n(X_i))^2. \tag{5.1}$$

To understand the limiting behavior of the test statistic, we will use the following heuristic arguments. Let us fix n and assume that $f_* \in \mathcal{G}$. Since $\tilde{g}_n$ minimizes $\sum_{i=1}^n (\tilde{Y}_i - f_*(X_i))^2$ over $f \in \mathcal{G}$, it follows that

$$\frac{1}{n} \sum_{i=1}^n (\tilde{Y}_i - \tilde{g}_n(X_i))^2 \le \frac{1}{n} \sum_{i=1}^n (\tilde{Y}_i - f_*(X_i))^2,$$

or equivalently,

$$\mathrm{TS} = \frac{m}{n} \sum_{i=1}^n (\tilde{Y}_i - \tilde{g}_n(X_i))^2 \le \frac{m}{n} \sum_{i=1}^n (\tilde{Y}_i - f_*(X_i))^2. \tag{5.2}$$

Using the fact that $\tilde{Y}_i = \sum_{j=1}^m (f_*(X_i) + \varepsilon_{ij})/m$ for $1 \le i \le n$, (5.2) can be rewritten as

$$\mathrm{TS} \le \frac{m}{n} \sum_{i=1}^n (\tilde{Y}_i - f_*(X_i))^2 = \frac{m}{n} \sum_{i=1}^n \left(\sum_{j=1}^m \varepsilon_{ij}/m \right)^2. \tag{5.3}$$

For each $i \in \{1, \cdots, n\}$, the weak law of large numbers ensures

$$\frac{1}{\sqrt{m}} \sum_{j=1}^m \varepsilon_{ij} \Rightarrow N(0, \sigma^2) \tag{5.4}$$

as $m \to \infty$, where $N(0, \sigma^2)$ is the normal random variable with a mean of 0 and a variance of σ^2. By the continuous mapping theorem together with (5.3) and

(5.4), we obtain

$$\text{TS} \le \frac{m}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{m}\varepsilon_{ij}/m\right)^2 \Rightarrow \frac{1}{n}\left(N_1(0,\sigma^2)+\cdots+N_n(0,\sigma^2)\right) \stackrel{D}{=} (\sigma^2/n)\chi^2_n,$$

where the $N_i(0,\sigma^2)$'s are independent normal random variables with a mean of 0 and a variance of σ^2, $\Rightarrow$ denotes convergence in distribution, and $A \stackrel{D}{=} B$ means that A and B have the same distribution.

On the other hand, let us fix n and assume that $f_* \notin \mathcal{G}$. Then, $\sum_{i=1}^n(\tilde{Y}_i - \tilde{g}_n(X_i))^2/n$ is expected to converge to $\sum_{i=1}^n(f_*(X_i)-g_*(X_i))^2/n$ as $m\to\infty$, which is a positive number. Hence, $\text{TS} = (m/n)\sum_{i=1}^n(\tilde{Y}_i-\tilde{g}_n(X_i))^2$ is expected to go to infinity as $m\to\infty$.

From the above arguments, it is intuitively acceptable to conclude that $f_*\in\mathcal{G}$ with a confidence level of at least $1-\gamma$ if the test statistic is less than the $100(1-\gamma)$–th percentile of $(\sigma^2/n)\chi^2_n$.

For a test procedure to be meaningful, it should fail to reject H_0 when $f_*\notin\mathcal{G}$. As our next proposition suggests, our test procedure has the desired property. Proposition 5.1 states that the Type I error of the test procedure converges to 0 as $n\to\infty$ for m sufficiently large.

To describe the behavior of the test procedure, we need some assumptions.

B1. $((X_i,Y_{ij}) : 1\le i\le n, 1\le j\le m)$ is a sequence of iid $(0,1)^d\times\mathbb{R}$–valued random vectors satisfying

$$Y_{ij} = f_*(X_i)+\varepsilon_{ij} \text{ for } i\ge 1 \text{ and } j\ge 1.$$

B2. $X_1, X_2, \cdots$ have a common positive density function $\tau : (0,1)^d\to\mathbb{R}$.
B3. There exists a positive constant τ_* such that $\tau(x)\le\tau_*$ for $x\in(0,1)^d$.
B4. (i) $\mathbb{E}(\varepsilon_i|X_i)=0$ and (ii) $\mathbb{E}(\varepsilon_i^2|X_i^2)=\sigma^2<\infty$ for $i\ge 1$.
B5. $\mathbb{E}f_*(X_1)^2<\infty$.
B6. There exists a solution to (3.2) that is continuous. Also, f_* is continuous.

Proposition 5.1. *Under B1–B6, P1–P3, there exists a constant M such that $m\ge M$ implies*

$$\mathbb{P}(\textit{Fail to reject } H_0 \textit{ for all but finitely many } n \mid f_*\notin\mathcal{G}) = 1. \tag{5.5}$$

If $\mathcal{G}=\mathcal{G}_m$ or $\mathcal{G}=\mathcal{G}_c$, (5.5) holds under B1–B6.

The proof of Proposition 5.1 is provided in Appendix A.

5.2. Simulation Studies

In this section, we examine the performance of the test procedure through simulations. We compute the proportion of times H_0 is rejected with nine different regression functions, $f_1, f_2, f_3, f_4, f_5, f_6, f_7, f_8$ and f_9, for different n values.

We define $f_1 : (0,1) \to \mathbb{R}$, $f_2 : (0,1) \to \mathbb{R}$, $f_3 : (0,1) \to \mathbb{R}$, $f_4 : (0,1) \to \mathbb{R}$, $f_5 : (0,1)^2 \to \mathbb{R}$, $f_6 : (0,1)^2 \to \mathbb{R}$, $f_7 : (0,1)^2 \to \mathbb{R}$, $f_8 : (0,1)^2 \to \mathbb{R}$, and $f_9 : (0,1)^2 \to \mathbb{R}$ by

$$\begin{aligned}
f_1(x) &= -\exp(-x) \\
f_2(x) &= -1.5(x-0.7)^2 \\
f_3(x) &= 5(x-0.2)^2 \\
f_4(x) &= 3(x-0.2)(x-0.5)(x-0.7) \\
f_5(x(1),x(2)) &= x(1)^2 + x(2)^2 \\
f_6(x(1),x(2)) &= x(1)^2 - x(2)^2 \\
f_7(x(1),x(2)) &= 1 - \exp(-(x(1)-0.5)^2 - (x(2)-0.5)^2) \\
f_8(x(1),x(2)) &= (x(1)+x(2)-0.5)^3 \\
&\quad - \exp(-50(x(1)-0.25)^2 + (x(2)-0.25)^2)
\end{aligned}$$

for $x \in (0,1)$ and $(x(1),x(2)) \in (0,1)^2$. On the other hand, $f_9(x)$ is the price of a European call option on a non–dividend–paying stock when $x \in (0,1)$ is the volatility of the underlying stock price. It is assumed that the strike price of this stock option is 1.3, the risk-free annual interest rate is 0.03, the current price of the underlying stock is 1, and the time to maturity is 1 year.

It should be noted that f_1 is monotone, f_2, f_3, and f_4 are not monotone, f_5 is convex, and f_6, f_7, and f_8 are not convex. The non-convexity of f_9 can be easily checked; see, for example, page 295 in [47]. Thus, we applied the test procedure with $\mathcal{G} = \mathcal{G}_m$ to the cases where $f_* = f_1, f_2, f_3$, and f_4, respectively. We also applied the test procedure with $\mathcal{G} = \mathcal{G}_c$ to the cases where $f_* = f_5, f_6, f_7, f_8$, and f_9, respectively.

When $f_* = f_1, f_2, f_3$ or f_4, we chose $X_i = i/n - 1/(2n)$ for $1 \le i \le n$ and generated $Y_{ij} = f_2(X_i) + N_{ij}(0, 0.1^2)$ for $1 \le i \le n$ and $1 \le j \le m$ with $m = 20$, where the $N_{ij}(0, 0.1^2)$'s are iid normal random variables with a mean of 0 and a variance of 0.1^2. We next computed $\tilde{Y}_i = \sum_{j=1}^m Y_{ij}/m$ for $1 \le i \le n$, $\tilde{g}_n$ as the solution to (3.1) with $\mathcal{G} = \mathcal{G}_m$ and the Y_i's replaced by the $\tilde{Y}_i$'s, and the test statistic in (5.1). When computing $\tilde{g}_n$, the quadratic programming formulation of (3.1) is solved through CVX, which is a software package designed to solve convex programs; see [33] for details. The test procedure is conducted with $\gamma = 0.05$, reaching the conclusion of whether H_0 should be rejected or not. We repeated this procedure 100 times independently. Using these 100 trials, we computed the 95% confidence intervals of the proportion of times H_0 is rejected. Table 1 reports these confidence intervals for a variety of n values.

When $f_* = f_5, f_6, f_7$ or f_8, we chose $\{X_1, \cdots, X_n\}$ as $\{(v/\sqrt{n} - 1/(2\sqrt{n}), w/\sqrt{n} - 1/(2\sqrt{n})) : 1 \le v, w \le \sqrt{n}\}$ and generated $Y_{ij} = f_2(X_i) + N_{ij}(0, 0.2^2)$ for $1 \le i \le n$ and $1 \le j \le m$ with $m = 10$, where the $N_{ij}(0, 0.2^2)$'s are iid normal random variables with a mean of 0 and a variance of 0.2^2. We next computed $\tilde{Y}_i = \sum_{j=1}^m Y_{ij}/m$ for $1 \le i \le n$, $\tilde{g}_n$ as the solution to (3.1) with $\mathcal{G} = \mathcal{G}_c$ and the Y_i's replaced by the $\tilde{Y}_i$'s, and the test statistic in (5.1). When computing $\tilde{g}_n$, the quadratic programming formulation of (3.1) is solved through

TABLE 1
The 95% confidence interval of the proportion of times rejecting H_0 when f_* is f_1, f_2, f_3, and f_4, respectively.

n	$f_* = f_1$	$f_* = f_2$	$f_* = f_3$	$f_* = f_4$
10	1.00 ± 0.00	0.72 ± 0.09	0.75 ± 0.08	0.92 ± 0.05
20	1.00 ± 0.00	0.50 ± 0.10	0.36 ± 0.09	0.82 ± 0.08
30	1.00 ± 0.00	0.23 ± 0.08	0.07 ± 0.05	0.57 ± 0.10
100	1.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.19 ± 0.08
200	1.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.02 ± 0.03

TABLE 2
The 95% confidence interval of the proportion of times rejecting H_0 when f_* is f_5, f_6, f_7, and f_8, respectively.

n	$f_* = f_5$	$f_* = f_6$	$f_* = f_7$	$f_* = f_8$
4	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
16	1.00 ± 0.00	0.56 ± 0.10	0.57 ± 0.10	1.00 ± 0.00
36	1.00 ± 0.00	0.07 ± 0.05	0.03 ± 0.03	0.00 ± 0.00
64	1.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
100	1.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00

CVX. The test procedure is conducted with $\gamma = 0.05$. We repeated this procedure 100 times independently. Using these 100 trials, we computed the 95% confidence intervals of the proportion of times H_0 is rejected. Table 2 reports these confidence intervals for a variety of n values.

When $f_* = f_9$, we chose $X_i = i/n - 1/(2n)$ for $1 \le i \le n$. For each fixed i, we simulated the underlying stock price up to one year from now using the geometric Brownian motion with a drift of 0.03 and a volatility of X_i. (The underlying stock price at the current time is assumed to be 1.) From the jth replication of this simulation ($1 \le j \le m$ with $m = 10$), we obtained S_{ij}, the price of the underlying stock one year from the current time. The price of the call option, Y_{ij}, is then obtained from $\exp(-0.03)\max(0, S_{ij} - 1.3)$. We then computed $\tilde{Y}_i = \sum_{j=1}^{m} Y_{ij}/m$ for $1 \le i \le n$, $\tilde{g}_n$ as the solution to (3.1) with $\mathcal{G} = \mathcal{G}_c$ and the Y_i's replaced by the $\tilde{Y}_i$'s, and the test statistic in (5.1). The test procedure is conducted with $\gamma = 0.05$. We repeated this procedure 100 times independently. Using these 100 trials, we computed the 95% confidence intervals of the proportion of times H_0 is rejected. Table 3 reports these confidence intervals for a variety of n values.

Tables 1, 2, and 3 show that the proportion of times H_0 is rejected gets closer to 0 as n increases for non-monotone or non-convex test functions, i.e. $f_2, f_3, f_4, f_6, f_7, f_8$, and f_9. This illustrates that the test procedure successfully detects the non–monotonicity or non-convexity of the test functions for large n, as Proposition 5.1 suggests. Furthermore, Tables 1 and 2 indicate that the test procedure also detects monotonicity of f_1 and convexity of f_5 for a variety of n values.

TABLE 3
The 95% confidence interval of the proportion of times rejecting H_0 when $f_* = f_9$.

n	$f_* = f_9$
10	0.62 ± 0.10
20	0.44 ± 0.10
30	0.14 ± 0.07
40	0.05 ± 0.04
50	0.03 ± 0.03

Appendix A: Proofs

This Appendix contains the proofs of Theorem 3.1, Corollaries 3.1–3.4, Theorem 4.1, Corollaries 4.1–4.2, and Proposition 5.1.

For notational convenience, $\mathbb{E}\psi(\hat{g}_n(X))$ for any measurable function $\psi : \mathbb{R} \to \mathbb{R}$ will denote, in the rest of this paper, $\int \psi(\hat{g}_n(x))dH(x)$, where H is the distribution function of X.

Proof of Theorem 3.1. The proof of Theorem 3.1 consists of 10 steps.

Step 1: We use the fact that $\hat{g}_n$ is a minimizer of φ_n and that $g_* \in \mathcal{G}$ to obtain

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{g}_n(X_i))^2 \le \frac{1}{n}\sum_{i=1}^{n}(Y_i - g_*)^2,$$

or equivalently,

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{g}_n(X_i) - g_*(X_i))^2 \le \frac{2}{n}\sum_{i=1}^{n}(Y_i - g_*(X_i))(\hat{g}_n(X_i) - g_*(X_i)). \tag{A.1}$$

Step 2: We will use A1, A4, and A5 to show that there exists a constant β such that

$$\frac{1}{n}\sum_{i=1}^{n}\hat{g}_n(X_i)^2 \le \beta \quad \text{a.s. for } n \text{ sufficiently large.} \tag{A.2}$$

To fill in the details, we apply the Cauchy-Schwarz inequality to the right-hand side of (A.1) to get

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{g}_n(X_i) - g_*(X_i))^2 \le 2\sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - g_*(X_i))^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{g}_n(X_i) - g_*(X_i))^2}$$

and

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{g}_n(X_i) - g_*(X_i))^2 \le \frac{4}{n}\sum_{i=1}^{n}(Y_i - g_*(X_i))^2. \tag{A.3}$$

We then apply $(a+b)^2 \le 2a^2+2b^2$ for $a,b\in\mathbb{R}$ to get

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n \hat{g}_n(X_i)^2 &\le \frac{2}{n}\sum_{i=1}^n(\hat{g}_n(X_i)-g_*(X_i))^2+\frac{2}{n}\sum_{i=1}^n g_*(X_i)^2\\ &\le \frac{8}{n}\sum_{i=1}^n(Y_i-g_*(X_i))^2+\frac{2}{n}\sum_{i=1}^n g_*(X_i)^2 \text{ by (A.3)}\\ &\le 8\mathbb{E}(Y-g_*(X))^2+\mathbb{E}(g_*(X)^2)+1 \triangleq \beta\end{aligned}$$

a.s. for n sufficiently large by the strong law of large numbers.

Step 3: We will use Step 2, A2, A3, and P3 to show

$$\mathbb{E}(\hat{g}_n(X)^2)\le \beta+1 \quad \text{a.s. for } n \text{ sufficiently lage.} \tag{A.4}$$

To fill in the details, we define the truncated value as follows: For any $c\ge 0$ and $x\ge 0$, the truncated value of x, denoted by $T_c(x)$, is defined by

$$T_c(x)=\begin{cases} x, & \text{if } x\ge c\\ 0, & \text{otherwize.}\end{cases} \tag{A.5}$$

Note that $T_c(x^2)=\{T_{\sqrt{c}}(x)\}^2$ for $x\ge 0$, $T_c(x)\le T_c(y)$ for $0\le x\le y$, and $T(x+y)\le T(x)+T(y)$ for $x,y\ge 0$. Together with the fact that $|x+y|\le |x|+|y|$ for $x,y\in\mathbb{R}$, we obtain

$$T_c(x+y)^2 \le T_c(x)^2+T_c(y^2)+2T_{\sqrt{c}}|x|\cdot T_{\sqrt{c}}|y| \tag{A.6}$$

$$-T_c(x+y)^2 \le -T_c(x)^2-T_c(y^2)+2T_{\sqrt{c}}|x|\cdot T_{\sqrt{c}}|y|. \tag{A.7}$$

We will show that for any $c>0$,

$$\mathbb{E}(T_c(\hat{g}_n(X)^2))\le \beta+1 \tag{A.8}$$

for n sufficiently large. If (A.8) is proven, letting $c\uparrow\infty$ for each n and applying the monotone convergence theorem will yield (A.4). It remains to show (A.8). Let $c>0$ and $\epsilon>0$ be given. We take $\delta>0$ small enough so that

$$c\mathbb{P}(X\in A_\delta)\le \epsilon, \tag{A.9}$$

where $A_\delta=(0,1)^d-[\delta,1-\delta]^d$. By P3, there exist a constant $r\triangleq r(\epsilon,\delta)$ and $\{h_1,\cdots,h_r\}\subset\tilde{\mathcal{H}}(\delta)$ such that (3.3) holds. Without loss of generality, we may assume that $\{h_1,\cdots,h_r\}$ is also an ϵ-net of $(\tilde{\mathcal{H}}(\delta),d_2^\delta)$.

First, we note

$$\begin{aligned}&\mathbb{E}T_c(\hat{g}_n(X)^2)=\mathbb{E}T_c(\hat{g}_n(X)^2)I(X\in A_\delta)+\mathbb{E}T_c(\hat{g}_n(X)^2)I(X\in[\delta,1-\delta]^d)\\ &\le c^2\mathbb{P}(X\in A_\delta)+\mathbb{E}T_c(\hat{g}_n(X)^2)I(X\in[\delta,1-\delta]^d)\\ &\le \epsilon+\mathbb{E}T_c(\hat{g}_n(X)^2)I(X\in[\delta,1-\delta]^d) \quad \text{by (A.9).}\end{aligned} \tag{A.10}$$

Next, let $B_\delta = [\delta, 1-\delta]^d$ and note that for each $j \in \{1, \cdots, r\}$,

$$
\begin{aligned}
&\mathbb{E}T_c(\hat{g}_n(X)^2)I(X \in B_\delta) - \frac{1}{n}\sum_{i=1}^n T_c(\hat{g}_n(X_i)^2)I(X_i \in B_\delta) \\
&= \mathbb{E}T_c(\hat{g}_n(X) - h_j(X) + h_j(X))^2 I(X \in B_\delta) \\
&\quad - \frac{1}{n}\sum_{i=1}^n T_c(\hat{g}_n(X_i) - h_j(X_i) + h_j(X_i))^2 I(X_i \in B_\delta) \\
&\le \mathbb{E}T_c(\hat{g}_n(X) - h_j(X))^2 I(X \in B_\delta) + \mathbb{E}T_c h_j(X)^2 I(X \in B_\delta) \\
&\quad + 2\mathbb{E}T_{\sqrt{c}}|\hat{g}_n(X) - h_j(X)| \cdot T_{\sqrt{c}}|h_j(X)| I(X \in B_\delta) \\
&\quad - \frac{1}{n}\sum_{i=1}^n T_c(\hat{g}_n(X_i) - h_j(X_i))^2 I(X_i \in B_\delta) - \frac{1}{n}\sum_{i=1}^n T_c h_j(X_i)^2 I(X_i \in B_\delta) \\
&\quad + \frac{2}{n}\sum_{i=1}^n T_{\sqrt{c}}|\hat{g}_n(X_i) - h_j(X_i)| \cdot T_{\sqrt{c}}|h_j(X_i)| I(X_i \in B_\delta) \text{ by (A.6) and (A.7)} \\
&\le \mathbb{E}T_c(\hat{g}_n(X) - h_j(X))^2 I(X \in B_\delta) + \mathbb{E}T_c h_j(X)^2 I(X \in B_\delta) \\
&\quad + 2\sqrt{\mathbb{E}T_c(\hat{g}_n(X) - h_j(X))^2 I(X \in B_\delta)} \cdot \sqrt{\mathbb{E}T_c(h_j(X)^2) I(X \in B_\delta)} \\
&\quad - \frac{1}{n}\sum_{i=1}^n T_c(\hat{g}_n(X_i) - h_j(X_i))^2 I(X_i \in B_\delta) - \frac{1}{n}\sum_{i=1}^n T_c h_j(X_i)^2 I(X_i \in B_\delta) \\
&\quad + 2\sqrt{\frac{1}{n}\sum_{i=1}^n T_c(\hat{g}_n(X_i) - h_j(X_i))^2 I(X_i \in B_\delta)} \\
&\quad \cdot \sqrt{\frac{1}{n}\sum_{i=1}^n T_c(h_j(X_i)^2) I(X_i \in B_\delta)} \text{ by the Cauchy-Schwarz inequality} \\
&\le \max_{1 \le j \le r} \left\{ \mathbb{E}T_c h_j(X)^2 I(X \in B_\delta) - \frac{1}{n}\sum_{i=1}^n T_c h_j(X_i)^2 I(X_i \in B_\delta) \right\} \\
&\quad + \mathbb{E}T_c(\hat{g}_n(X) - h_j(X))^2 I(X \in B_\delta) \\
&\quad + 2\sqrt{\mathbb{E}T_c(\hat{g}_n(X) - h_j(X))^2 I(X \in B_\delta)} \cdot \sqrt{\mathbb{E}T_c(h_j(X)^2) I(X \in B_\delta)} \\
&\quad - \frac{1}{n}\sum_{i=1}^n T_c(\hat{g}_n(X_i) - h_j(X_i))^2 I(X_i \in B_\delta) \\
&\quad + 2\sqrt{\frac{1}{n}\sum_{i=1}^n T_c(\hat{g}_n(X_i) - h_j(X_i))^2 I(X_i \in B_\delta)} \sqrt{\frac{1}{n}\sum_{i=1}^n T_c(h_j(X_i)^2) I(X_i \in B_\delta)},
\end{aligned}
$$

so,

$$
\mathbb{E}T_c(\hat{g}_n(X)^2)I(X \in B_\delta) - \frac{1}{n}\sum_{i=1}^n T_c(\hat{g}_n(X_i)^2)I(X_i \in B_\delta)
$$

$$
\begin{aligned}
&\leq \max_{1\leq j\leq r}\left\{\mathbb{E}T_c h_j(X)^2 I(X\in B_\delta) - \frac{1}{n}\sum_{i=1}^n T_c h_j(X_i)^2 I(X_i\in B_\delta)\right\}\\
&\quad + \min_{1\leq j\leq r}\mathbb{E}T_c(\hat{g}_n(X)-h_j(X))^2 I(X\in B_\delta)\\
&\quad + 2\min_{1\leq j\leq r}\sqrt{\mathbb{E}T_c(\hat{g}_n(X)-h_j(X))^2 I(X\in B_\delta)}\cdot\sqrt{\mathbb{E}T_c(h_j(X)^2)I(X\in B_\delta)}\\
&\quad + 2\min_{1\leq j\leq r}\sqrt{\frac{1}{n}\sum_{i=1}^n T_c(\hat{g}_n(X_i)-h_j(X_i))^2 I(X_i\in B_\delta)}\\
&\quad \cdot\sqrt{\frac{1}{n}\sum_{i=1}^n T_c(h_j(X_i)^2)I(X_i\in B_\delta)}\\
&\quad \leq \epsilon
\end{aligned}
\tag{A.11}
$$

a.s. for n sufficiently large by the strong law of large numbers and P3. So, combining (A.10), (A.11), and the fact that

$$
\frac{1}{n}\sum_{i=1}^n T_c(\hat{g}_n(X_i)^2)I(X_i\in B_\delta) \leq \frac{1}{n}\sum_{i=1}^n \hat{g}_n(X_i)^2
$$

yields

$$
\mathbb{E}T_c(\hat{g}_n(X)^2) \leq 2\epsilon + \frac{1}{n}\sum_{i=1}^n \hat{g}_n(X_i)^2
$$

a.s. for n sufficiently large. Using (A.2), we can conclude

$$
\limsup_{n\to\infty}\mathbb{E}T_c(\hat{g}_n(X)^2) \leq 2\epsilon+\beta \quad \text{a.s.},
$$

which implies (A.8).

Step 4: For $\delta>0$, let $A_\delta = (0,1)^d\setminus[\delta,1-\delta]^d$. We will show that for each $\epsilon>0$, there exists $\delta>0$ small enough so that

$$
\frac{1}{n}\sum_{i=1}^n (Y_i-g_*(X_i))(\hat{g}_n(X_i)-g_*(X_i))I(X_i\in A_\delta)\leq\epsilon \text{ and} \tag{A.12}
$$

$$
\frac{1}{n}\sum_{i=1}^n (f_*(X_i)-g_*(X_i))(\hat{g}_n(X_i)-g_*(X_i))I(X_i\in A_\delta)\leq\epsilon \tag{A.13}
$$

a.s. for n sufficiently large.

To fill in the details, let $\epsilon>0$ be given. We apply the Cauchy-Schwarz inequality to obtain

$$
\frac{1}{n}\sum_{i=1}^n (Y_i-g_*(X_i))(\hat{g}_n(X_i)-g_*(X_i))I(X_i\in A_\delta)
$$

$$\begin{aligned}
&\leq \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - g_*(X_i))^2 I(X_i \in A_\delta)}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{g}_n(X_i) - g_*(X_i))^2} \\
&\leq \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - g_*(X_i))^2 I(X_i \in A_\delta)}\sqrt{\frac{2}{n}\sum_{i=1}^{n}(Y_i - g_*(X_i))^2} \text{ by (A.3)} \\
&\leq \sqrt{\mathbb{E}(Y - g_*(X))^2 I(X \in A_\delta) + \epsilon}\sqrt{2\mathbb{E}(Y - g_*(X))^2 + \epsilon}
\end{aligned}$$

a.s. for n sufficiently large. We then take δ small enough so that $\mathbb{E}((Y - g_*(X))^2 I(X \in A_\delta)) < \epsilon$, completing the proof of (A.12). Similarly, (A.13) follows.

Step 5: For $\delta > 0$, let $B_\delta = [\delta, 1-\delta]^d$. We will use P3 to prove that for any $\delta > 0$,

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}(f_*(X_i) - g_*(X_i))(\hat{g}_n(X_i) - g_*(X_i))I(X_i \in B_\delta) \\
&- \mathbb{E}(f_*(X) - g_*(X))(\hat{g}_n(X) - g_*(X))I(X \in B_\delta) \to 0
\end{aligned} \tag{A.14}$$

a.s. as $n \to \infty$.

To fill in the details, let $\epsilon > 0$ and $\delta > 0$ be given. By P3, there exist a constant $r \triangleq r(\epsilon, \delta)$ and $\{h_1, \cdots, h_r\} \subset \tilde{\mathcal{H}}(\delta)$ satisfying (3.5). Without loss of generality, we may assume that $\{h_1, \cdots, h_r\}$ is also an ϵ-net of $(\tilde{\mathcal{H}}(\delta), d_2^\delta)$. Then, for each $j \in \{1, \cdots, r\}$,

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}(f_*(X_i) - g_*(X_i))(\hat{g}_n(X_i) - g_*(X_i))I(X_i \in B_\delta) \\
&- \mathbb{E}(f_*(X) - g_*(X))(\hat{g}_n(X) - g_*(X))I(X \in B_\delta) \\
&= \frac{1}{n}\sum_{i=1}^{n}(f_*(X_i) - g_*(X_i))(\hat{g}_n(X_i) - h_j(X_i))I(X_i \in B_\delta) \\
&\quad + \frac{1}{n}\sum_{i=1}^{n}(f_*(X_i) - g_*(X_i))(h_j(X_i) - g_*(X_i))I(X_i \in B_\delta) \\
&\quad - \mathbb{E}(f_*(X) - g_*(X))(\hat{g}_n(X) - h_j(X))I(X \in B_\delta) \\
&\quad - \mathbb{E}(f_*(X) - g_*(X))(h_j(X) - g_*(X))I(X \in B_\delta), \\
&\leq \max_{1 \leq j \leq r}\left\{\frac{1}{n}\sum_{i=1}^{n}(f_*(X_i) - g_*(X_i))(h_j(X_i) - g_*(X_i))I(X_i \in B_\delta)\right. \\
&\quad \left. - \mathbb{E}(f_*(X) - g_*(X))(h_j(X) - g_*(X))I(X \in B_\delta)\right\} \\
&\quad + \sqrt{\frac{1}{n}\sum_{i=1}^{n}(f_*(X_i) - g_*(X_i))^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{g}_n(X_i) - h_j(X_i))^2 I(X_i \in B_\delta)} \\
&\quad + \sqrt{\mathbb{E}(f_*(X) - g_*(X))^2}\sqrt{\mathbb{E}(\hat{g}_n(X) - h_j(X))^2 I(X \in B_\delta)}
\end{aligned}$$

by the Cauchy-Schwarz inequality, so

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}(f_*(X_i)-g_*(X_i))(\hat{g}_n(X_i)-g_*(X_i))I(X_i\in B_\delta)\\
&-\mathbb{E}(f_*(X)-g_*(X))(\hat{g}_n(X)-g_*(X))I(X\in B_\delta)\\
&\le \max_{1\le j\le r}\left\{\frac{1}{n}\sum_{i=1}^{n}(f_*(X_i)-g_*(X_i))(h_j(X_i)-g_*(X_i))I(X_i\in B_\delta)\right.\\
&\quad\left.-\mathbb{E}(f_*(X)-g_*(X))(h_j(X)-g_*(X))I(X\in B_\delta)\right\}\\
&\quad+\min_{1\le j\le r}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(f_*(X_i)-g_*(X_i))^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{g}_n(X_i)-h_j(X_i))^2I(X_i\in B_\delta)}\\
&\quad+\min_{1\le j\le r}\sqrt{\mathbb{E}(f_*(X)-g_*(X))^2}\sqrt{\mathbb{E}(\hat{g}_n(X)-h_j(X))^2I(X\in B_\delta)}\\
&\le \epsilon+2\epsilon(\sqrt{\mathbb{E}(f_*(X)-g_*(X))^2}+\epsilon)
\end{aligned}$$

a.s. for n sufficiently large due to (3.5) and the strong law of large numbers, proving (A.14).

Step 6: We will use Step 5 to prove

$$\limsup_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}(f_*(X_i)-g_*(X_i))(\hat{g}_n(X_i)-g_*(X_i))\le 0\quad\text{a.s.}\tag{A.15}$$

To fill in the details, let $\epsilon>0$ be given. We then take $\delta>0$ small enough so that (A.13) holds and

$$\sqrt{\mathbb{E}(f_*(X)-g_*(X))^2I(X\in A_\delta)}\sqrt{2\beta+3+2\mathbb{E}g_*(X)^2}\le\epsilon,\tag{A.16}$$

where $A_\delta=(0,1)^d\setminus[\delta,1-\delta]^d$. Let $B_\delta=[\delta,1-\delta]^d$ and observe that

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}(f_*(X_i)-g_*(X_i))(\hat{g}_n(X_i)-g_*(X_i))-\mathbb{E}(f_*(X)-g_*(X))(\hat{g}_n(X)-g_*(X))\\
&=\frac{1}{n}\sum_{i=1}^{n}(f_*(X_i)-g_*(X_i))(\hat{g}_n(X_i)-g_*(X_i))I(X_i\in A_\delta)\\
&\quad+\frac{1}{n}\sum_{i=1}^{n}(f_*(X_i)-g_*(X_i))(\hat{g}_n(X_i)-g_*(X_i))I(X_i\in B_\delta)\\
&\quad-\mathbb{E}(f_*(X)-g_*(X))(\hat{g}_n(X)-g_*(X))I(X\in A_\delta)\\
&\quad-\mathbb{E}(f_*(X)-g_*(X))(\hat{g}_n(X)-g_*(X))I(X\in B_\delta)\\
&\le\frac{1}{n}\sum_{i=1}^{n}(f_*(X_i)-g_*(X_i))(\hat{g}_n(X_i)-g_*(X_i))I(X_i\in A_\delta)
\end{aligned}$$

$$+\frac{1}{n}\sum_{i=1}^{n}(f_*(X_i)-g_*(X_i))(\hat{g}_n(X_i)-g_*(X_i))I(X_i\in B_\delta)$$
$$-\mathbb{E}(f_*(X)-g_*(X))(\hat{g}_n(X)-g_*(X))I(X\in B_\delta)$$
$$+\sqrt{\mathbb{E}(f_*(X)-g_*(X))^2I(X\in A_\delta)}\sqrt{\mathbb{E}(\hat{g}_n(X)-g_*(X))^2}\le\epsilon$$

a.s. for n sufficiently large by (A.13), (A.14), (A.4), and (A.16). Since

$$\mathbb{E}(f_*(X)-g_*(X))(\hat{g}_n(X)-g_*(X))\le 0$$

by (3.2), (A.15) follows.

Step 7: We will use Step 3 and P3 to prove that

$$\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(\hat{g}_n(X_i)-g_*(X_i))\to 0 \text{ as } n\to\infty \text{ a.s.} \tag{A.17}$$

Let $\epsilon>0$ be given. We then take $\delta>0$ small enough so that

$$\sqrt{2\mathbb{E}\varepsilon^2I(X\in A_\delta)}\sqrt{2\beta+3+2\mathbb{E}g_*(X)^2}\le\epsilon, \tag{A.18}$$

where $A_\delta=(0,1)^2\setminus[\delta,1-\delta]^d$.

First, note that,

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(\hat{g}_n(X_i)-g_*(X_i))I(X_i\in A_\delta)\right|\\
&\le\sqrt{\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2I(X_i\in A_\delta)}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{g}_n(X_i)-g_*(X_i))^2}\\
&\le\sqrt{\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2I(X_i\in A_\delta)}\sqrt{\frac{2}{n}\sum_{i=1}^{n}\hat{g}_n(X_i)^2+\frac{2}{n}\sum_{i=1}^{n}g_*(X_i)^2}\\
&\le\sqrt{2\mathbb{E}\varepsilon^2I(X\in A_\delta)}\sqrt{2\beta+3+2\mathbb{E}g_*(X)^2}\le\epsilon
\end{aligned}$$

a.s. for n sufficiently large by Step 3, (A.18), and the strong law of large numbers. So, it follows that

$$\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(\hat{g}_n(X_i)-g_*(X_i))I(X_i\in A_\delta)\to 0 \text{ as } n\to\infty \text{ a.s.} \tag{A.19}$$

Next, let $B_\delta=[\delta,1-\delta]^d$. By P3, there exist a constant $r\triangleq r(\epsilon,\delta)$ and $\{h_1,\cdots,h_r\}\subset\tilde{\mathcal{H}}(\delta)$ satisfying (3.5). Observe that for each $j\in\{1,\cdots,r\}$,

$$\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(\hat{g}_n(X_i)-g_*(X_i))I(X_i\in B_\delta)\right|\le\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(\hat{g}_n(X_i)-h_j(X_i))I(X_i\in B_\delta)\right|$$

$$+\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(h_j(X_i)-g_*(X_i))I(X_i\in B_\delta)\right|$$
$$\le\sqrt{\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{g}_n(X_i)-h_j(X_i))^2I(X_i\in B_\delta)}$$
$$+\max_{1\le j\le r}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(h_j(X_i)-g_*(X_i))I(X_i\in B_\delta)\right|$$

by the Cauchy-Schwarz inequality, so

$$\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(\hat{g}_n(X_i)-g_*(X_i))I(X_i\in B_\delta)\right|$$
$$\le\min_{1\le j\le r}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{g}_n(X_i)-h_j(X_i))^2I(X_i\in B_\delta)}$$
$$+\max_{1\le j\le r}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(h_j(X_i)-g_*(X_i))I(X_i\in B_\delta)\right|$$
$$\le\epsilon\sqrt{\mathbb{E}(\varepsilon^2)+1}+\epsilon$$

a.s. for n sufficiently large by (3.5) and the strong law of large numbers. Hence,

$$\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(\hat{g}_n(X_i)-g_*(X_i))I(X_i\in B_\delta)\to 0 \text{ as } n\to\infty \text{ a.s.} \tag{A.20}$$

Combining (A.19) and (A.20) yields (A.17).

Step 8: We combine Steps 6 and 7 to establish (3.6). First, note that the combination of Steps 6 and 7 yields

$$\limsup_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}(Y_i-g_*(X_i))(\hat{g}_n(X_i)-g_*(X_i))\le 0 \text{ a.s.}$$

Using (A.1), we conclude that

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{g}_n(X_i)-g_*(X_i))^2\to 0 \text{ as } n\to\infty \text{ a.s.}$$

Step 9: We will use Step 8 and P3 to establish (3.7). To fill in the details, we will first prove that

$$\mathbb{E}T_c(\hat{g}_n(X)-g_*(X))^2\to 0 \text{ as } n\to\infty \tag{A.21}$$

for each $c>0$, where $T_c(\cdot)$ is defined as in (A.5). Once (A.21) is proven, letting $c\uparrow\infty$ for each n will prove (3.7). It remains to show (A.21). Let $\epsilon>0$ and $c>0$ be given. We then take $\delta>0$ small enough so that

$$c\mathbb{P}(X\in A_\delta)\le\epsilon,$$

where $A_\delta = (0,1)^d \setminus [\delta, 1-\delta]^d$. It follows that

$$\mathbb{E}T_c(\hat{g}_n(X) - g_*(X))^2 I(X \in A_\delta) \le \epsilon \tag{A.22}$$

Let $B_\delta = [\delta, 1-\delta]$. By P3, there exist a constant $r \triangleq r(\epsilon, \delta)$ and $\{h_1, \cdots, h_r\} \subset \tilde{\mathcal{H}}(\delta)$ satisfying (3.5). Without loss of generality, we may assume $\{h_1, \cdots, h_r\} \subset \tilde{\mathcal{H}}(\delta)$ is an ϵ-net of $(\tilde{\mathcal{H}}(\delta), d_2^\delta)$. Then, for each $j \in \{1, \cdots, r\}$,

$$\begin{aligned}
&\mathbb{E}T_c(\hat{g}_n(X) - g_*(X))^2 I(X \in B_\delta) - \frac{1}{n}\sum_{i=1}^n T_c(\hat{g}_n(X_i) - g_*(X_i))^2 I(X_i \in B_\delta) \\
&= \mathbb{E}T_c(\hat{g}_n(X) - h_j(X_i) + h_j(X_i) - g_*(X))^2 I(X \in B_\delta) \\
&\quad - \frac{1}{n}\sum_{i=1}^n T_c(\hat{g}_n(X_i) - h_j(X_i) + h_j(X_i) - g_*(X_i))^2 I(X_i \in B_\delta) \\
&\le \mathbb{E}T_c(\hat{g}_n(X) - h_j(X))^2 I(X \in B_\delta) + \mathbb{E}T_c(h_j(X) - g_*(X))^2 I(X \in B_\delta) \\
&\quad + 2\mathbb{E}T_{\sqrt{c}}|\hat{g}_n(X) - h_j(X)| \cdot T_{\sqrt{c}}|h_j(X) - g_*(X)| I(X \in B_\delta) \\
&\quad - \frac{1}{n}\sum_{i=1}^n T_c(\hat{g}_n(X_i) - h_j(X_i))^2 I(X_i \in B_\delta) \\
&\quad - \frac{1}{n}\sum_{i=1}^n T_c(h_j(X_i) - g_*(X_i))^2 I(X_i \in B_\delta) \\
&\quad + \frac{2}{n}\sum_{i=1}^n T_{\sqrt{c}}|\hat{g}_n(X_i) - h_j(X_i)| \cdot T_{\sqrt{c}}|h_j(X_i) - g_*(X_i)| I(X_i \in B_\delta) \\
&\quad \text{by (A.6) and (A.7)} \\
&\le \max_{1 \le j \le r} \Bigg\{ \mathbb{E}T_c(h_j(X) - g_*(X))^2 I(X \in B_\delta) \\
&\quad - \frac{1}{n}\sum_{i=1}^n T_c(h_j(X_i) - g_*(X_i))^2 I(X_i \in B_\delta) \Bigg\} \\
&\quad + \mathbb{E}T_c(\hat{g}_n(X) - h_j(X_i))^2 I(X \in B_\delta) \\
&\quad - \frac{1}{n}\sum_{i=1}^n T_c(\hat{g}_n(X_i) - h_j(X_i))^2 I(X_i \in B_\delta) \\
&\quad + 2\sqrt{\mathbb{E}T_c(\hat{g}_n(X) - h_j(X))^2 I(X \in B_\delta)} \cdot \sqrt{\mathbb{E}T_c(h_j(X) - g_*(X))^2 I(X \in B_\delta)} \\
&\quad + 2\sqrt{\frac{1}{n}\sum_{i=1}^n T_c(\hat{g}_n(X_i) - h_j(X_i))^2 I(X_i \in B_\delta)} \\
&\quad \cdot \sqrt{\frac{1}{n}\sum_{i=1}^n T_c(h_j(X_i) - g_*(X_i))^2 I(X_i \in B_\delta)}
\end{aligned}$$

by the Cauchy-Schwarz inequality, so

$$\begin{aligned}
&\mathbb{E}T_c(\hat{g}_n(X) - g_*(X))^2 I(X \in B_\delta) - \frac{1}{n}\sum_{i=1}^{n} T_c(\hat{g}_n(X_i) - g_*(X_i))^2 I(X_i \in B_\delta) \\
&\leq \max_{1\leq j\leq r} \Big\{ \mathbb{E}T_c(h_j(X) - g_*(X))^2 I(X \in B_\delta) \\
&\quad - \frac{1}{n}\sum_{i=1}^{n} T_c(h_j(X_i) - g_*(X_i))^2 I(X_i \in B_\delta) \Big\} \\
&\quad + \min_{1\leq j\leq r} \mathbb{E}T_c(\hat{g}_n(X) - h_j(X_i))^2 I(X \in B_\delta) \\
&+ 2 \min_{1\leq j\leq r} \sqrt{\mathbb{E}T_c(\hat{g}_n(X) - h_j(X))^2 I(X \in B_\delta)} \\
&\quad \cdot \sqrt{\mathbb{E}T_c(h_j(X) - g_*(X))^2 I(X \in B_\delta)} \\
&\quad + 2 \min_{1\leq j\leq r} \sqrt{\frac{1}{n}\sum_{i=1}^{n} T_c(\hat{g}_n(X_i) - h_j(X_i))^2 I(X_i \in B_\delta)} \\
&\quad \cdot \sqrt{\frac{1}{n}\sum_{i=1}^{n} T_c(h_j(X_i) - g_*(X_i))^2 I(X_i \in B_\delta)} \\
&\leq \epsilon
\end{aligned} \tag{A.23}$$

a.s. for n sufficiently large by (3.5) and the strong law of large numbers. The combination of (A.22), (A.23), and Step 8 proves (A.21), and hence, (3.7) is established.

Step 10: We will use Step 8 and P4 to establish (3.8). Let $\delta > 0$ be given. By P4, there exists a constant $\beta(\delta)$ such that

$$\begin{aligned}
&\sup_{\hat{g}_n \in \hat{\mathcal{G}}_n} |\hat{g}_n(x) - \hat{g}_n(y)| \leq \beta(\delta)\|x - y\| \text{ and} \\
&\sup_{g_* \in \mathcal{G}_*} |g_*(x) - g_*(y)| \leq \beta(\delta)\|x - y\|
\end{aligned}$$

a.s. for $x \in [\delta, 1-\delta]^d$ and n sufficiently large. We next find a finite number of hyperrectangles $S_1, \cdots, S_l$ with non-empty interior, satisfying $\cup_{j=1}^{l} S_j = [\delta, 1-\delta]^d$ and $\|x - y\| \leq \epsilon$ for $x, y \in S_j$ and $1 \leq j \leq l$.

For each $x \in S_j$ and $X_i \in S_j$,

$$\begin{aligned}
|\hat{g}_n(x) - g_*(x)| &\leq |\hat{g}_n(x) - \hat{g}_n(X_i)| + |\hat{g}_n(X_i) - g_*(X_i)| + |g_*(X_i) - g_*(x)| \\
&\leq \beta(\delta)\epsilon + |\hat{g}_n(X_i) - g_*(X_i)| + \beta(\delta)\epsilon,
\end{aligned}$$

so

$$\sup_{x \in S_j} |\hat{g}_n(x) - g_*(x)|$$

$$
\begin{aligned}
&\leq 2\beta(\delta)\epsilon + \frac{\sum_{i=1}^n |\hat{g}_n(X_i) - g_*(X_i)| I(X_i \in S_j)}{\sum_{i=1}^n I(X_i \in S_j)} \\
&\leq 2\beta(\delta)\epsilon + \frac{1}{n}\sum_{i=1}^n |\hat{g}_n(X_i) - g_*(X_i)| I(X_i \in S_j) \cdot \frac{n}{\sum_{i=1}^n I(X_i \in S_j)} \\
&\leq 2\beta(\delta)\epsilon + \sqrt{\frac{1}{n}\sum_{i=1}^n (\hat{g}_n(X_i) - g_*(X_i))^2} \cdot \frac{n}{\sum_{i=1}^n I(X_i \in S_j)} \\
&\quad \text{by the Cauchy-Schwarz inequality} \\
&\leq 2\beta(\delta)\epsilon + \epsilon
\end{aligned}
$$

a.s. for n sufficiently large by Step 8 and A2. Since there are finitely many S_j's,

$$\sup_{x \in [\delta, 1-\delta]^d} |\hat{g}_n(x) - g_*(x)| \leq 2\beta(\delta)\epsilon + \epsilon$$

a.s. for n sufficiently large, establishing (3.8). □

Proof of Corollary 3.1. The proof of Corollary 3.1 consists of six steps. In Steps 1–5, we will prove that A1–A3 imply P1–P3. In Step 6, we will consider the case where $d = 1$ and f_* is continuous and will prove that g_* exists uniquely and (3.8) holds.

Step 1: We establish P1. The only non-trivial part is that $\mathcal{G}_m$ is closed in L^2. Suppose that $(h_l : l \geq 1)$ is a Cauchy sequence in $\mathcal{G}_m$. Since L^2 is complete, there exists $h_\infty \in L^2$ satisfying $\|h_l - h_\infty\|_2 \to 0$ as $l \to \infty$. Furthermore, there exists a subsequence $(l_k : k \geq 1)$ and $\Omega \subset (0,1)^d$ such that $\lim_{k\to\infty} h_{l_k}(x) = h_\infty(x) < \infty$ for $x \in \Omega$ and Ω is a set of Lebesgue measure 1. We define $\tilde{h}_\infty : (0,1)^d \to \mathbb{R}$ by

$$\tilde{h}_\infty(x) = \sup\{h_\infty(z) : z \leq x, x \in \Omega\}$$

for $x \in (0,1)^d$. Then, $\tilde{h}_\infty$ is well–defined, monotone, and coincides with h_∞ on Ω. Therefore, $\|\tilde{h}_\infty - h_l\|_2 \to 0$ as $l \to \infty$. So, $\mathcal{G}_m$ is closed.

Step 2: We establish P2. Let $S_n = \{(g(X_1), \cdots g(X_n)) : g \in \mathcal{G}_m\}$. Then S_n is a nonempty and convex subset of $\mathbb{R}^n$. To see why S_n is closed, let $\{g_l : l \geq 1\}$ be a subset of $\mathcal{G}_m$ satisfying $|g_l(X_i) - \overline{g}_i| \to 0$ as $l \to \infty$ for $1 \leq i \leq n$ and some $(\overline{g}_1, \cdots, \overline{g}_n) \in \mathbb{R}^n$. Next, we define $\tilde{g}_\infty : (0,1)^d \to \mathbb{R}$ by

$$\tilde{g}_\infty(x) = \begin{cases} \sup\{\overline{g}_i : X_i \leq x, 1 \leq i \leq n\}, & \text{if there is } X_i \text{ satisfying } X_i \leq x \\ \min\{\overline{g}_i : 1 \leq i \leq n\}, & \text{otherwise} \end{cases}$$

for $x \in (0,1)^d$. Then $\tilde{g}_\infty \in \mathcal{G}_m$ and

$$\|(g_l(X_1), \cdots, g_l(X_n)) - (\tilde{g}_\infty(X_1), \cdots, \tilde{g}_\infty(X_n))\| \to 0$$

as $l \to \infty$.

Since φ_n is a strictly convex function on S_n, and S_n is nonempty, closed, convex, there exists a minimizer $(\tilde{g}_1, \cdots, \tilde{g}_n)$ of φ_n over S_n. We define $\tilde{g} : (0,1)^d \to \mathbb{R}$ by

$$\tilde{g}(x) = \begin{cases} \sup\{\tilde{g}_i : X_i \leq x, 1 \leq i \leq n\}, & \text{if there is } X_i \text{ satisfying } X_i \leq x \\ \min\{\tilde{g}_i : 1 \leq i \leq n\}, & \text{otherwise} \end{cases}$$

for $x \in (0,1)^d$. Then $\tilde{g}$ minimizes $\varphi_n(g)$ over $g \in \mathcal{G}$.

Step 3: We will prove that for any subset A of $(0,1)^d$ with nonempty interior, there exists a constant c_A such that

$$\inf_{X_i \in A} |\hat{g}_n(X_i) - g_*(X_i)| \le c_A \tag{A.24}$$

a.s. for n sufficiently large.

To fill in the details, note

$$\begin{aligned}
&\inf_{X_i \in A} |\hat{g}_n(X_i) - g_*(X_i)| \\
&\quad \le \frac{\sum_{i=1}^n |\hat{g}_n(X_i) - g_*(X_i)| I(X_i \in A)}{\sum_{i=1}^n I(X_i \in A)} \\
&\quad \le \frac{n}{\sum_{i=1}^n I(X_i \in A)} \cdot \frac{\sum_{i=1}^n |\hat{g}_n(X_i) - g_*(X_i)| I(X_i \in A)}{n} \\
&\quad \le \frac{n}{\sum_{i=1}^n I(X_i \in A)} \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{g}_n(X_i) - g_*(X_i))^2} \\
&\qquad \text{by the Cauchy-Schwarz inequality} \\
&\quad \le \frac{n}{\sum_{i=1}^n I(X_i \in A)} \sqrt{\frac{2}{n} \sum_{i=1}^n \hat{g}_n(X_i)^2 + \frac{2}{n} \sum_{i=1}^n g_*(X_i)^2} \\
&\quad \le \frac{2}{\mathbb{P}(X \in A)} \sqrt{2\beta + 3 + 2\mathbb{E}(g_*(X)^2)} \triangleq c_A
\end{aligned}$$

by (A.2) a.s. for n sufficiently large, proving (A.24).

Step 4: We will use Step 3 and the fact that g_* and $\hat{g}_n$ are monotone to prove that for any $\delta > 0$, there exists a constant $\beta_0(\delta)$ such that

$$\sup_{x \in [\delta, 1-\delta]^d} |\hat{g}_n(x)| \le \beta_0(\delta) \tag{A.25}$$

a.s. for n sufficiently large.

To fill in the details, let $\delta > 0$ be given. (A.24) implies that there exists a constant c_δ, $(X_{i_k} : k \ge 1)$, and $(X_{j_k} : j \ge 1)$ such that

$$\begin{aligned}
&X_{i_k} \in (3\delta/4, \delta)^d, && X_{j_k} \in (1-\delta, 1-3\delta/4)^d, \\
&g_*(X_{i_k}) - c_\delta \le \hat{g}_n(X_{i_k}), \text{ and} && \hat{g}_n(X_{j_k}) \le g_*(X_{j_k}) + c_\delta
\end{aligned}$$

for $k \ge 1$. So, for any $x \in [\delta, 1-\delta]^d$, the monotonicity of $\hat{g}_n$ implies

$$g_*(X_{i_k}) - c_\delta \le \hat{g}_n(X_{i_k}) \le \hat{g}_n(x) \le \hat{g}_n(x_{j_k}) \le g_*(X_{j_k}) + c_\delta,$$

and

$$g_*(X_{i_k}) - c_\delta \le \hat{g}_n(x) \le g_*(X_{i_k}) + c_\delta.$$

For any $X_i \in (\delta/2, 3\delta/4)^d$ and $X_j \in (1-3\delta/4, 1-\delta/2)^d$, the monotonicity of g_* implies

$$g_*(X_i) - c_\delta \le \hat{g}_n(x) \le g_*(X_j) + c_\delta.$$

Therefore,

$$\max_{X_i \in (\delta/2, 3\delta/4)^d} g_*(X_i) - c_\delta \le \hat{g}_n(x) \le \min_{X_j \in (1-3\delta/4, 1-\delta/2)^d} g_*(X_j) + c_\delta. \quad \text{(A.26)}$$

The first inequality of (A.26) implies

$$\begin{aligned}
&\hat{g}_n(x) \qquad \text{(A.27)}\\
&\ge \frac{\sum_{i=1}^n g_*(X_i) I(X_i \in (\delta/2, 3\delta/4)^d)}{\sum_{i=1}^n I(X_i \in (\delta/2, 3\delta/4)^d)} - c_\delta \\
&\ge \frac{n}{\sum_{i=1}^n I(X_i \in (\delta/2, 3\delta/4)^d)} \frac{1}{n} \sum_{i=1}^n g_*(X_i) I(X_i \in (\delta/2, 3\delta/4)^d) - c_\delta \\
&\ge -\frac{n}{\sum_{i=1}^n I(X_i \in (\delta/2, 3\delta/4)^d)} \sqrt{\frac{1}{n} \sum_{i=1}^n g_*(X_i)^2 I(X_i \in (\delta/2, 3\delta/4)^d)} - c_\delta \\
&\quad \text{by the Cauchy-Schwarz inequality} \\
&\ge -\frac{4}{\mathbb{P}(X \in (\delta/2, 3\delta/4)^d)} \mathbb{E} g_*(X)^2 I(X \in (\delta/2, 3\delta/4)^d) - c_\delta \qquad \text{(A.28)}
\end{aligned}$$

a.s. for n sufficiently large by the strong law of large numbers. Similarly, the second inequality of (A.26) implies

$$\begin{aligned}
\hat{g}_n(x) \le \frac{4}{\mathbb{P}(X \in (1-3\delta/4, 1-\delta/2)^d)} \\
\cdot \mathbb{E} g_*(X)^2 I(X \in (1-3\delta/4, 1-\delta/2)^d) + c_\delta \qquad \text{(A.29)}
\end{aligned}$$

a.s. for n sufficiently large. The combination of (A.28) and (A.29) proves (A.25).

Step 5: We will use Step 4 to establish P3. Let $\delta > 0$ be given and define $\mathcal{H}(\delta)$ and $\tilde{\mathcal{H}}(\delta)$ by

$$\begin{aligned}
\mathcal{H}(\delta) &= \{h \in \mathcal{G}_m : |h(x)| \le \beta_0(\delta) \text{ for } x \in [\delta, 1-\delta]^d\} \\
\tilde{\mathcal{H}}(\delta) &= \{\tilde{h} : [\delta, 1-\delta]^d \to \mathbb{R} : \text{There exists } h \in \mathcal{H}(\delta) \text{ such that} \\
&\qquad \tilde{h}(x) = h(x) \text{ for } x \in [\delta, 1-\delta]^d\}.
\end{aligned}$$

We first note that $N(\epsilon, \tilde{\mathcal{H}}(\delta), d_2^\delta) < \infty$ for any $\epsilon > 0$; see, for example, Theorem 1.1 on page 1752 of [30]. To establish P3(ii), let $\epsilon > 0$ be given. Without loss of generality, we may assume that $h(x) \ge 0$ for any $h \in \tilde{\mathcal{H}}(\delta)$. We divide $[0,1]^d$ into hyperrectangles $[a^k, b^k]$, $1 \le k \le K$, with the side length $1/l$, where $l = \lceil (1-2\delta)/\epsilon^2 \rceil$. The volume of each hyperrectangle is less than $\epsilon^{2d}/(1-2\delta)^d$. Note that K is of the order $1/\epsilon^{2d}$. Let us consider the number of hyperrectangles that intersect with a face of $[\delta, 1-\delta]^d$, i.e., a set of the form

$$\{(x_1, \cdots, x_d) \in [\delta, 1-\delta]^d : x_j = \delta\} \text{ or } \{(x_1, \cdots, x_d) \in [\delta, 1-\delta]^d : x_j = 1-\delta\}$$

for some $j \in \{1, \cdots, d\}$. Then, the number of such hyperrectangles is of the order $1/\epsilon^{2(d-1)}$.

For $1 \le k \le K$, we define

$$H_k = \frac{1}{n} \sum_{i=1}^{n} I(X_i \in [a^k, b^k]).$$

By the strong law of large numbers and A3,

$$H_k \le \mathbb{E} I(X \in [a^k, b^k]) + \epsilon^{2d} \le \tau_* \epsilon^{2d} / (1 - 2\delta)^d + \epsilon^{2d}$$

a.s. for n sufficiently large. Note that n only depends on a^k and b^k, $1 \le k \le K$.

For each $h \in \tilde{\mathcal{H}}(\delta)$ and $k \in \{1, \cdots, K\}$, we define

$$g_k = \begin{cases} \frac{\sum_{i=1}^n g(X_i) I(X_i \in [a^k, b^k])}{\sum_{i=1}^n I(X_i \in [a^k, b^k])}, & \text{if } \sum_{i=1}^n I(X_i \in [a^k, b^k]) > 0, \\ 0, & \text{otherwise} \end{cases}$$

and let $\overline{g}_k = \lfloor g_k / \epsilon \rfloor$. Then, for $g \in \tilde{\mathcal{H}}(\delta)$,

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^{n} (g(X_i) - \epsilon \overline{g}_k I(X_i \in [a^k, b^k]))^2 &\le \epsilon^2 H_k + (g(b^k)^2 - g(a^k)^2) H_k \\ &\le \epsilon^2 H_k + (g(b^k)^2 - g(a^k)^2)(1 + \tau_* / (1 - 2\delta)^d) \epsilon^{2d} \end{aligned}$$

a.s. for n sufficiently large, and hence,

$$d_n^\delta \left(g, \sum_{k=1}^{K} \epsilon \overline{g}_k I(x \in [a^k, b^k]) \right)^2 \le c(\delta) \epsilon^2$$

for some constant $c(\delta)$ a.s. for n sufficiently large. Since $\overline{g}_k$, $1 \le k \le K$, is an integer between 0 and $\lceil \beta_0(\delta) \rceil$ and K is an integer that is of the order $1/\epsilon^{2d}$, there are finitely many functions of the form $\sum_{k=1}^K \epsilon \overline{g}_k I(x \in [a^k, b^k])$, proving (3.5).

Step 6: Since P1–P3 are established, (3.6) and (3.7) hold. We next suppose that $d = 1$ and f_* is continuous. Then g_* is continuous (see Lemma 1 on page 252 of [36]), and hence, exists uniquely. We will use (3.6), the fact that g_* is continuous, the fact that $\hat{g}_n$ is monotone in order to establish (3.8).

Let $\delta > 0$ and $\epsilon > 0$ be given. Without loss of generality, we may assume $\delta < \epsilon$. Since g_* is continuous over $[\delta/2, 1 - \delta/2]^d$, there exists $\lambda(\delta)$ such that

$$|g_*(x) - g_*(y)| \le \lambda(\delta) \|x - y\|^d$$

for $x, y \in [\delta/2, 1 - \delta/2]^d$. We next divide $[\delta/2, 1 - \delta/2]^d$ into hyperrectangles $[a^k, b^k]$, $1 \le k \le K$, with side-length less than ϵ satisfying $\cup_{k=1}^l [a^k, b^k] = [\delta, 1 - \delta]^d$ for some $l \le K$. Note that for any $x \in [a^i, b^i] \subset [\delta, 1 - \delta]^d$, there are hyperrectangles $[a^i, b^i]$ and $[a^k, b^k]$ so that they share vertices with $[a^i, b^i]$, any

point in $[a^j, b^j]$ is less than or equal to any point in $[a^i, b^i]$, and any point in $[a^k, b^k]$ is greater than or equal to any point in $[a^i, b^i]$. In other words, there exist $j \triangleq j(i)$ and $k \triangleq k(i)$ such that

$$X_s \le x \le X_t \text{ for } X_s \in [a^j, b^j] \text{ and } X_t \in [a^k, b^k]$$
$$\|x - X_s\| \le 2\sqrt{d}\epsilon \text{ for } X_s \in [a^j, b^j] \text{ or } X_s \in [a^k, b^k].$$

Then, for $x \in [a^i, b^i] \subset [\delta, 1-\delta]^d$, $X_s \in [a^j, b^j]$, and $X_t \in [a^k, b^k]$,

$$\hat{g}_n(X_s) - g_*(x) \le \hat{g}_n(x) - g_*(x) \le \hat{g}_n(X_t) - g_*(x),$$

and hence,

$$\hat{g}_n(X_s) - g_*(X_s) + g_*(X_s) - g_*(x) \le \hat{g}_n(x) - g_*(x) \le \hat{g}_n(X_t) - g_*(X_t) + g_*(X_t) - g_*(x)$$

so,

$$|\hat{g}_n(x) - g_*(x)| \le \max\{|\hat{g}_n(X_s) - g_*(X_s)| + |g_*(X_s) - g_*(x)|, |\hat{g}_n(X_t) - g_*(X_t)| + |g_*(X_t) - g_*(x)|\}.$$

Therefore, for each $X_s \in [a^j, b^j]$, and $X_t \in [a^k, b^k]$,

$$\begin{aligned} &|\hat{g}_n(x) - g_*(x)| \\ &\le \max\{|\hat{g}_n(X_s) - g_*(X_s)| + \max_{y \in [a^j, b^j]} |g_*(y) - g_*(x)|, \\ &\qquad |\hat{g}_n(X_t) - g_*(X_t)| + \max_{y \in [a^k, b^k]} |g_*(y) - g_*(x)|\} \\ &\le |\hat{g}_n(X_s) - g_*(X_s)| + |\hat{g}_n(X_t) - g_*(X_t)| + 2\sqrt{d}\lambda(\delta)\epsilon, \end{aligned}$$

so

$$|\hat{g}_n(x) - g_*(x)| \le \min_{X_s \in [a^j, b^j]} |\hat{g}_n(X_s) - g_*(X_s)| + \min_{X_t \in [a^k, b^k]} |\hat{g}_n(X_t) - g_*(X_t)| + 2\sqrt{d}\lambda(\delta)\epsilon,$$

and

$$\begin{aligned} \sup_{x \in [a^i, b^i]} |\hat{g}_n(x) - g_*(x)| &\le \min_{X_s \in [a^j, b^j]} |\hat{g}_n(X_s) - g_*(X_s)| \\ &\quad + \min_{X_t \in [a^k, b^k]} |\hat{g}_n(X_t) - g_*(X_t)| + 2\sqrt{d}\lambda(\delta)\epsilon \\ &\le \frac{\sum_{l=1}^n |\hat{g}_n(X_l) - g_*(X_l)| I(X_l \in [a^j, b^j])}{\sum_{l=1}^n I(X_l \in [a^j, b^j])} \\ &\quad + \frac{\sum_{l=1}^n |\hat{g}_n(X_l) - g_*(X_l)| I(X_l \in [a^k, b^k])}{\sum_{l=1}^n I(X_l \in [a^k, b^k])} + 2\sqrt{d}\lambda(\delta)\epsilon \end{aligned}$$

$$
\begin{aligned}
&\leq \frac{n}{\sum_{l=1}^{n} I(X_l \in [a^j, b^j])} \cdot \frac{1}{n} \sum_{l=1}^{n} |\hat{g}_n(X_l) - g_*(X_l)| I(X_l \in [a^j, b^j]) \\
&\quad + \frac{n}{\sum_{l=1}^{n} I(X_l \in [a^k, b^k])} \cdot \frac{1}{n} \sum_{l=1}^{n} |\hat{g}_n(X_l) - g_*(X_l)| I(X_l \in [a^k, b^k]) \\
&\quad + 2\sqrt{d}\lambda(\delta)\epsilon \\
&\leq \frac{n}{\sum_{l=1}^{n} I(X_l \in [a^j, b^j])} \sqrt{\frac{1}{n} \sum_{l=1}^{n} (\hat{g}_n(X_l) - g_*(X_l))^2 I(X_l \in [a^j, b^j])} \\
&\quad + \frac{n}{\sum_{l=1}^{n} I(X_l \in [a^k, b^k])} \sqrt{\frac{1}{n} \sum_{l=1}^{n} (\hat{g}_n(X_l) - g_*(X_l))^2 I(X_l \in [a^k, b^k])} \\
&\quad + 2\sqrt{d}\lambda(\delta)\epsilon \\
&\leq \epsilon + 2\sqrt{d}\lambda(\delta)\epsilon \qquad \text{(A.30)}
\end{aligned}
$$

a.s. for n sufficiently large by (3.6).

Since there are finitely many hyperrectangles $[a^i, b^i]$, we conclude

$$\sup_{x \in [\delta, 1-\delta]^d} |\hat{g}_n(x) - g_*(x)| \to 0$$

as $n \to \infty$ a.s. □

Proof of Corollary 3.2. We will prove that A1–A3 imply P1–P4.

Step 1: We establish P1 and the uniqueness of the solution to (3.2). When proving P1, the only non-trivial part is that $\mathcal{G}_c$ is closed in L^2. Suppose $(h_l : l \geq 1)$ is a Cauchy sequence in $\mathcal{G}_c$. Since L^2 is complete, there exists $h_\infty \in L^2$ satisfying $\|h_l - h_\infty\|_2 \to 0$ as $l \to \infty$. Furthermore, there exists a subsequence $(l_k : k \geq 1)$ and $\Omega \subset (0,1)^d$ such that $\lim_{k\to\infty} h_{l_k}(x) = h_\infty(x)$ for $x \in \Omega$ and Ω is a set of Lebesgue measure one. We next define $\tilde{h}_\infty : (0,1)^d \to \mathbb{R}$ by

$$\tilde{h}_\infty(x) = \sup\{h_\infty(z) + \eta_z^T(x - z) : z \in \Omega, \eta_z \in \partial h_\infty(z)\},$$

where $\partial h_\infty(z)$ is the set of the subgradients of h_∞ at $z \in \Omega$. Then $\tilde{h}_\infty$ is convex, coincides with h_∞ on Ω, and hence, $\|h_l - \tilde{h}_\infty\|_2 \to 0$ as $l \to \infty$. So, $\mathcal{G}_c$ is closed.

Now, we turn to the uniqueness of the solution to (3.2). Since $\mathcal{G}_c$ is a nonempty, closed convex subset of L^2, the projection theorem implies that there exists a solution g_* to (3.2) and the solution is unique up to a set of measure zero. In other words, if g_* and $\tilde{g}_*$ are two solutions to (3.2), then $g_*(x) = \tilde{g}_*(x)$ almost everywhere. But, both g_* and $\tilde{g}_*$ are convex over $(0,1)^d$, and hence, continuous, so they must agree at $x \in (0,1)^d$.

Step 2: We establish P2. To fill in the details, let $S_n = \{(g(X_1), \cdots, g(X_n)) : g \in \mathcal{G}_c\}$. Then, S_n is a non-empty, closed, and convex subset of $\mathbb{R}^n$; see, for example, Lemma 2.3 on page 1638 of [65]. Since $\varphi_n(z_1, \cdots, z_n) = \sum_{i=1}^{n}(Y_i - z_i)^2/n$ is a strictly convex function over $(z_1, \cdots, z_n) \in S_n$, there exists a unique

minimizer $(\tilde{z}_1, \cdots, \tilde{z}_n)$ of φ_n over S_n. Since $(\tilde{z}_1, \cdots, \tilde{z}_n) \in S_n$, there exists $\tilde{g}_n \in \mathcal{G}_c$ such that $\tilde{g}_n(X_i) = \tilde{z}_i$ for $1 \le i \le n$. Then, $\tilde{g}_n$ becomes a solution to (3.1).

Step 3: We establish P3 and P4. To fill in the details, we notice that for any $\delta > 0$, there exist constants $\beta_0(\delta)$ and $\beta_1(\delta)$ such that

$$\sup_{x \in [\delta, 1-\delta]^d} |\hat{g}_n(x)| \le \beta_0(\delta) \text{ and}$$
$$|\hat{g}_n(x) - \hat{g}_n(y)| \le \beta_1(\delta)\|x - y\| \tag{A.31}$$

for $x, y \in [\delta, 1-\delta]^d$ and n sufficiently large a.s.; see, for example, Steps 4 and 5 on pages 201 and 202 of [53]. So, if we let

$$\mathcal{H}(\delta) = \{h \in \mathcal{G}_c : |h(x)| \le \beta_0(\delta),\ |h(x) - h(y)| \le \beta_1(\delta)\|x - y\| \text{ for } x \in [\delta, 1-\delta]^d\}$$

and

$$\tilde{\mathcal{H}}(\delta) = \{\tilde{h} : [\delta, 1-\delta]^d \to \mathbb{R} : \text{There exists } h \in \mathcal{H}(\delta) \text{ such that } \tilde{h}(x) = h(x) \text{ for } x \in [\delta, 1-\delta]^d\},$$

then (3.4) is established. To establish P3(i) and P3(ii), note that $\tilde{\mathcal{H}}(\delta)$ is a subset of $\mathcal{A}$, where

$$\mathcal{A} = \{h : [\delta, 1-\delta]^d \to \mathbb{R} : h \text{ is convex, } |h(x)| \le \beta_0(\delta) \ |h(x) - h(y)| \le \beta_1(\delta)\|x - y\| \text{ for } x, y \in [\delta, 1-\delta]^d\},$$

and hence,

$$N(\epsilon, \tilde{\mathcal{H}}(\delta), d_2^\delta) \le N(\epsilon, \mathcal{A}, d_\infty^\delta) < \infty$$
$$N(\epsilon, \tilde{\mathcal{H}}(\delta), d_n^\delta) \le N(\epsilon, \mathcal{A}, d_\infty^\delta) < \infty$$

by Theorem 6 of [11].

We can establish P4 by using (A.31) and the fact that g_* is unique and continuous over $(0,1)^d$. □

Proof of Corollary 3.3. The proof is similar to the proof of Corollary 3.1. □

Proof of Corollary 3.4. The proof is similar to the proof of Corollary 3.2. □

Proof of Theorem 4.1. We will use (4.1), the fact that f_*, g_*, $g \in \mathcal{G}$ are bounded uniformly, and that fact that the ε_i's are subexponential to apply Theorem 3.4.1 of [74]. We start by observing that $\hat{g}_n$ minimizes

$$\frac{1}{n}\sum_{i=1}^n (Y_i - g(X_i))^2 = \frac{1}{n}\sum_{i=1}^n ((f_*(X_i) - g_*(X_i) + \varepsilon_i) + (g_*(X_i) - g(X_i)))^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}(f_*(X_i) - g_*(X_i) + \varepsilon_i)^2$$
$$- \frac{2}{n}\sum_{i=1}^{n}(f_*(X_i) - g_*(X_i) + \varepsilon_i)(g(X_i) - g_*(X_i)) + \frac{1}{n}\sum_{i=1}^{n}(g(X_i) - g_*(X_i))^2$$

over $g \in \mathcal{G}$, so it maximizes

$$\mathbb{M}_n(g) \triangleq \frac{2}{n}\sum_{i=1}^{n}(f_*(X_i) - g_*(X_i) + \varepsilon_i)(g(X_i) - g_*(X_i)) - \frac{1}{n}\sum_{i=1}^{n}(g(X_i) - g_*(X_i))^2.$$

For $g \in \mathcal{G}$, define $M_n(g)$ and $d(g, g_*)$ by

$$M_n(x) \triangleq 2\mathbb{E}(f_*(X) - g_*(X))(g(X) - g_*(X)) - \mathbb{E}(g(X) - g_*(X))^2$$

and

$$d(g, g_*) \triangleq \left\{\mathbb{E}(g(X) - g_*(X))^2\right\}^{1/2}.$$

Let

$$\mathcal{F} = \{2(f_*(X) - g_*(X) + \varepsilon)(g(X) - g_*(X)) - (g(X) - g_*(X))^2 : g \in \mathcal{G},\ \delta/2 < d(g, g_*) \le \delta\}.$$

For any $f \in \mathcal{F}$, define the Bernstein norm $\|\cdot\|_B$ by

$$\|f\|_B \triangleq \{2\mathbb{E}(\exp|f(X)| - 1 - |f(X)|)\}^{1/2}.$$

We next notice that Theorem 3.4.1 on pages 322 and 323 and Problem 3.4.2 on page 337 of [74] can be rewritten as follows:

Theorem A.1 (Due to Theorem 3.4.1 and Problem 3.4.2 of [74]). *Suppose that, for every n and $\delta \in (0, \infty)$,*

$$\sup\{M_n(g) - M_n(g_*) : \delta/2 < d(g, g_*) \le \delta,\ g \in \mathcal{G}\} \le -\delta^2/4 \tag{A.32}$$

and

$$\mathbb{E}\sup\{\sqrt{n}[\mathbb{M}_n(g) - M_n(g) - \delta^2/8]^+ : \delta/2 < d(g, g_*) \le \delta,\ g \in \mathcal{G}\} \lesssim \phi_n(\delta) \tag{A.33}$$

for functions ϕ_n such that $\delta \mapsto \phi_n(\delta)/\delta^\alpha$ is decreasing on $(0, \infty)$, for some $\alpha < 2$. Let r_n satisfy $r_n^2\phi_n(1/r_n) \lesssim \sqrt{n}$ for every n. If $\mathbb{M}_n(\hat{g}_n) \ge \mathbb{M}_n(g_*)$, then

$$r_n d(g, g_*) = \mathcal{O}_p(1)$$

as $n \to \infty$.

Also, Lemma 3.4.3 on page 324 and Problem 3.4.3 on page 337 of [74] can be rewritten as follows:

Lemma A.1 (Due to Lemma 3.4.3 and Problem 3.4.3 of [74]). *If $\mathcal{F}$ satisfies $\|f\|_B \le \delta$ for every $f \in \mathcal{F}$, then*

$$\begin{aligned}&\mathbb{E}\sup\{[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(f(X_i)-\mathbb{E}f(X))-\sqrt{n}(\delta^2/8\wedge\delta/3)]^+ : f\in\mathcal{F}\}\\ &\quad\le\quad \tilde{J}_{[\,]}(\delta,\mathcal{F},\|\cdot\|_B)(1+\tilde{J}_{[\,]}(\delta,\mathcal{F},\|\cdot\|_B)/(\delta^2\sqrt{n})),\end{aligned}\tag{A.34}$$

where

$$\tilde{J}_{[\,]}(\delta,\mathcal{F},\|\cdot\|_B)=\int_{(\delta^2/64)\wedge(\delta/24)\wedge\delta}^{\delta}\sqrt{1+\log N_{[\,]}(u,\mathcal{G},d_2)}du.$$

We next verify the two non-trivial conditions of Theorem A.1, which are (A.32) and (A.33). (A.32) is easily verified because

$$\begin{aligned}&M_n(x)-M_n(g_*)\\ &\quad=\quad 2\mathbb{E}(f_*(X)-g_*(X))(g(X)-g_*(X))-\mathbb{E}(g(X)-g_*(X))^2\\ &\quad\le\quad -\mathbb{E}(g(X)-g_*(X))^2 \text{ by (4.1)}\\ &\quad<\quad -\delta^2/4\end{aligned}$$

for g satisfying $d(g,g_*)>\delta/2$.

To verify (A.33) with ϕ_n specified in Theorem 4.1, we will use Lemma A.1. First, we will show that the Bernstein norm of an element of $\mathcal{F}$ is bounded by $\overline{c}_1\delta$ for some constant $\overline{c}_1$. To see why this is true, note that the squared Bernstein norm of a element of $\mathcal{F}$ is

$$\begin{aligned}&2\mathbb{E}(\exp|2(f_*(X)-g_*(X)+\varepsilon)(g(X)-g_*(X))-(g(X)-g_*(X))^2|\\ &-1-|2(f_*(X)-g_*(X)+\varepsilon)(g(X)-g_*(X))-g(X)-g_*(X))^2|)\\ &\quad=\quad 2\sum_{i\ge2}\frac{1}{i!}\mathbb{E}|2(f_*(X)-g_*(X)+\varepsilon)(g(X)-g_*(X))-(g(X)-g_*(X))^2|^i\\ &\qquad\text{by Taylor's Theorem}\\ &\quad=\quad 2\mathbb{E}|2(f_*(X)-g_*(X)+\varepsilon)(g(X)-g_*(X))-(g(X)-g_*(X))^2|^2\\ &\qquad\cdot\exp|2(f_*(X)-g_*(X)+\varepsilon)(g(X)-g_*(X))-(g(X)-g_*(X))^2|.\end{aligned}\tag{A.35}$$

Since

$$\begin{aligned}&\mathbb{E}|2(f_*(X)-g_*(X)+\varepsilon)(g(X)-g_*(X))-(g(X)-g_*(X))^2|\\ &=\mathbb{E}(2f_*(X)+2g_*(X)+2\varepsilon-4g(X))^2(g(X)-g_*(X))^2\\ &\le 128(A_*^2+B_*^2+\sigma^2)\mathbb{E}(g(X)-g_*(X))^2,\end{aligned}$$

f_*, g_*, and $g\in\mathcal{G}$ are uniformly bounded, and the ε_i's are subexponential, the squared Bernstein norm of an element in $\mathcal{F}$ in the left-hand side of (A.35) is bounded by $\overline{c}_1^2\mathbb{E}(g(X)-g_*(X))^2$ for some constant $\overline{c}_1$. So, the Bernstein norm of an element of $\mathcal{F}$ is bounded by

$$\overline{c}_1 d(g,g_*)\le\overline{c}_1\delta.\tag{A.36}$$

We will next show that

$$N_{[\,]}(\epsilon, \mathcal{F}, \|\cdot\|_B) \lesssim N_{[\,]}(\epsilon, \mathcal{G}, d_2). \tag{A.37}$$

To establish (A.37), let $[l_1, u_1], \cdots, [l_r, u_r]$ be ϵ–brackets that cover $\mathcal{G}$, i.e., $\mathcal{G} \subset \bigcup_{i=1}^{r}[l_j, u_j]$. Note that if $g \in \mathcal{G}$ satisfies $l_j \le g \le u_j$ for some $j \in \{1, \cdots, r\}$, its counterpart in $\mathcal{F}$ is

$$\begin{aligned} f_g(X) &\triangleq 2(f_*(X) - g_*(X) + \varepsilon)(g(X) - g_*(X)) - (g(X) - g_*(X))^2 \\ &= (2f_*(X) - g_*(X) + 2\varepsilon - g(X))(g(X) - g_*(X)) \\ &\triangleq v(X)w(X), \end{aligned}$$

where $v(X) = 2f_*(X) - g_*(X) + 2\varepsilon - g(X)$ and $w(X) = g(X) - g_*(X)$. $v(X)$ and $w(X)$ satisfy

$$\begin{aligned} v_m(X) \triangleq 2f_*(X) - g_*(X) + 2\varepsilon - u_j(X) \le v(X) \\ \le 2f_*(X) - g_*(X) + 2\varepsilon - l_j(X) \triangleq v_M(X) \end{aligned}$$

and

$$w_m(X) \triangleq l_j(X) - g_*(X) \le w(X) \le u_j(X) - g_*(X) \triangleq w_M(X).$$

Also, for any $v_1(X), w_1(X), v_2(X)$, and $w_2(X)$ satisfying

$$\begin{aligned} v_m(X) \le v_i(X) \le v_M(X) \quad i = 1, 2 \\ w_m(X) \le m_i(X) \le w_M(X) \quad i = 1, 2, \end{aligned}$$

we have

$$\begin{aligned} &|v_1(X)w_1(X) - v_2(X)w_2(X)| \\ &\le |v_1(X)||w_1(X) - w_2(X)| + |w_1(X)||v_1(X) - v_2(X)| \\ &\quad + |v_1(X) - v_2(X)||w_1(X) - w_2(X)| \\ &\le 4(A_* + B_* + |\varepsilon|)|u_j(X) - v_j(X)| + (u_j(X) - l_j(X))^2. \end{aligned}$$

Therefore, if we define $\tilde{l}_j$ and $\tilde{u}_j$ be

$$\begin{aligned} \tilde{u}_j(X) &= \max\{vw : v_m(X) \le v \le v_M(X), w_m(X) \le w \le w_M(X)\} \\ \tilde{l}_j(X) &= \tilde{u}_j(X) - 4(A_* + B_* + |\varepsilon|)|u_j(X) - l_j(X)| - (u_j(X) - l_j(X))^2, \end{aligned}$$

then $\tilde{l_j}(X) \le f_g(X) \le \tilde{u}_j(X)$ and

$$\begin{aligned} &\|\tilde{l}_j(X) - \tilde{u}_j(X)\|_B^2 \\ &= \|(l_j(X) - u_j(X))^2 + 4(A_* + B_* + |\varepsilon|)|u_j(X) - l_i(X)|\|_B^2 \\ &= \||u_j(X) - l_j(X)|(|u_j(X) - l_j(X)| + 4(A_* + B_* + |\varepsilon|))\|_B^2 \\ &\le 2\mathbb{E}(u_j(X) - l_j(X))^2(u_j(X) - l_j(X) + 4(A_* + B_* + |\varepsilon|))^2 \\ &\quad \cdot \exp|u_j(X) - l_j(X)|(|u_j(X) - l_j(X)| + 4(A_* + B_* + |\varepsilon|)) \end{aligned}$$

$$\leq \quad \overline{c}_2 \mathbb{E}(u_j(X) - l_j(X))^2$$

for some constant $\overline{c}_2$ since u_j and l_j are bounded by B_* and $|\varepsilon|$ is subexponential, proving (A.37). The combination of Theorem A.1, Lemma A.1, (A.36), and (A.37) proves Theorem 4.1. □

Proof of Corollary 4.1. We apply Theorem 4.1 to the case where $\mathcal{G} = \mathcal{G}_{m,B}$. By Theorem 1.1 of [30],

$$\log N_{[\,]}(\epsilon, \mathcal{G}_{m,B}, d_2) \lesssim \begin{cases} \epsilon^{-1}, & \text{if } d = 1 \\ \epsilon^{-2}(\log(1/\epsilon))^2, & \text{if } d = 2 \\ \epsilon^{-2(d-1)}, & \text{if } d > 2. \end{cases}$$

Thus, we obtain

$$\phi_n(\delta) = \begin{cases} \delta^{1/2}(1 + \delta^{1/2}/(\delta^2\sqrt{n})), & \text{if } d = 1 \\ (\log 1/\delta)^2(1 + (\log 1/\delta)^2/(\delta^2\sqrt{n})), & \text{if } d = 2 \\ \delta^{-2(d-2)}(1 + \delta^{-2(d-2)}/(\delta^2\sqrt{n})), & \text{if } d > 2, \end{cases}$$

and it can be verified that $\phi_n(\sqrt{a_n}) \lesssim \sqrt{n}a_n$, where a_n is given by (4.2). Therefore, Corollary 4.1 follows. □

Proof of Corollary 4.2. We apply Theorem 4.1 to the case where $\mathcal{G} = \mathcal{G}_{c,B}$. By Theorem 1.1 (ii) on page 567 of [31],

$$\log N_{[\,]}(\epsilon, \mathcal{G}_{c,B}, d_2) \lesssim \epsilon^{-d/2}.$$

Thus, we obtain

$$\phi_n(\delta) = \begin{cases} \delta^{1-d/4}(1 + \delta^{1-d/4}/(\delta^2\sqrt{n})), & \text{if } d < 4 \\ -\log\delta(1 - \log\delta/(\delta^2\sqrt{n})), & \text{if } d = 4 \\ \delta^{2-d/2}(1 + \delta^{2-d/2}/(\delta^2\sqrt{n})), & \text{if } d > 4, \end{cases}$$

and it can be verified that $\phi_n(\sqrt{b_n}) \lesssim \sqrt{n}b_n$, where b_n is given by (4.3). Therefore, Corollary 4.2 follows. □

Proof of Proposition 5.1. Throughout this proof, we will assume $f_* \notin \mathcal{G}$. The proof consists of four steps.

Step 1: We notice that B6 implies $\mathbb{E}(f_*(X_1) - g_*(X_1))^2 > 0$ when $f \notin \mathcal{G}$. Let g_* be a solution to (3.2) that is continuous. Suppose, on the contrary, that $\mathbb{E}(f_*(X_1) - g_*(X_1))^2 = 0$. Since both f_* and g_* are continuous, $f_*(x) = g_*(x)$ for $x \in (0,1)^d$, which contradicts $f_* \notin \mathcal{G}$. Let $\tilde{g}_*$ be any solution to (3.2). Since $\mathbb{E}(g_*(X_1) - \tilde{g}_*(X_1))^2 = 0$, we have $\mathbb{E}(f_*(X_1) - \tilde{g}_*(X_1))^2 = \mathbb{E}(f_*(X_1) - g_*(X_1))^2 > 0$.

Step 2: We next observe that B1–B5, P2, and P3 imply that

$$\limsup_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} (\tilde{g}_n(X_i) - g_*(X_i))(f_*(X_i) - g_*(X_i)) \leq 0$$

a.s. for each $m \geq 1$ by Step 6 in the Proof of Theorem 3.1.

Step 3: We will prove that there exists a positive constant $\tilde{c}$ such that

$$\frac{1}{n}\sum_{i=1}^{n}(\tilde{Y}_i - \tilde{g}_n(X_i))^2 \geq \tilde{c} \tag{A.38}$$

for n sufficiently large and each $m \geq 1$. To see why this is true, we note that

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}(\tilde{Y}_i - \tilde{g}_n(X_i))^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}((f_*(X_i) - g_*(X_i)) + (g_*(X_i) + \overline{\varepsilon}_i - \tilde{g}_n(X_i)))^2, \\
&\quad \text{where } \overline{\varepsilon}_i = \sum_{j=1}^{m}\varepsilon_{ij}/m \\
&= \frac{1}{n}\sum_{i=1}^{n}(f_*(X_i) - g_*(X_i))^2 + \frac{1}{n}\sum_{i=1}^{n}(g_*(X_i) + \overline{\varepsilon}_i - \tilde{g}_n(X_i))^2 \\
&\quad + \frac{2}{n}\sum_{i=1}^{n}\overline{\varepsilon}_i(f_*(X_i) - g_*(X_i)) \\
&\quad - \frac{2}{n}\sum_{i=1}^{n}(\tilde{g}_n(X_i) - g_*(X_i))(f_*(X_i) - g_*(X_i)) \\
&\geq \frac{1}{n}\sum_{i=1}^{n}(f_*(X_i) - g_*(X_i))^2 + \frac{2}{n}\sum_{i=1}^{n}\overline{\varepsilon}_i(f_*(X_i) - g_*(X_i)) \\
&\quad - \frac{2}{n}\sum_{i=1}^{n}(\tilde{g}_n(X_i) - g_*(X_i))(f_*(X_i) - g_*(X_i)) \\
&\geq (1/2)\mathbb{E}(f_*(X_1) - g_*(X_1))^2 \triangleq \tilde{c}
\end{aligned}$$

a.s. for n sufficiently large.

Step 4: We will next prove Proposition 5.1. We note that for each n,

$$z_{1-\gamma} < \sigma/\sqrt{1-\gamma}$$

because the Markov inequality states that for any $a > 0$,

$$\mathbb{P}(\sigma^2\chi_n^2/n > a) \leq \mathbb{E}(\sigma^2\chi_n^2/n)/a^2 = \sigma^2/a^2,$$

so taking $a = \sigma/\sqrt{1-\gamma}$ confirms that $\mathbb{P}(\sigma^2\chi_n^2/n > a) \leq 1-\gamma$ and $z_{1-\gamma} < a$. If we take m large enough so that $mc > \sigma/\sqrt{1-\gamma}$, then for n sufficiently large, (A.38) implies

$$\mathrm{TS} = \frac{m}{n}\sum_{i=1}^{n}(\tilde{Y}_i - \tilde{g}_n(X_i))^2 \geq mc > \sigma/\sqrt{1-\gamma} > z_{1-\gamma}$$

a.s. for n sufficiently large, and hence,

$$\mathbb{P}(\text{Fail to reject } H_0 \text{ for } n \text{ sufficiently large} \mid f_* \notin \mathcal{G})$$
$$= \mathbb{P}\left(\frac{m}{n}\sum_{i=1}^{n}(\tilde{Y}_n - \tilde{g}_n(X_i))^2 > z_{1-\gamma} \text{ for } n \text{ sufficiently large} \mid f_* \notin \mathcal{G}\right) = 1,$$

proving Proposition 5.1. □

Acknowledgements

The author wishes to express her thanks to the referees and associate editor for their suggestions and comments, which served to significantly improve the quality of the paper.

References

[1] Bacchetti, P. (1989). Additive isotonic models. *J. Amer. Statist. Assoc.* **84** 289–294. MR0999691

[2] Balabdaoui, F., Rufibach, K. and Wellner, J. A. (2009). Limit distribution theory for maximum likelihood estimation of a log–concave density. *Ann. Statist.* **37** 1299–1331. MR2509075

[3] Balabdaoui, F., Jankowski, H., Rufibach, K. and Pavlides, M. (2013). Asymptotics of the discrete log–concave maximum likelihood estimator and related applications. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **75** 769–790. MR3091658

[4] Banerjee, M. and Wellner, J. A. (2001). Likelihood ratio tests for monotone functions. *Ann. Statist.* **29** 1699–1731. MR1891743

[5] Baraud, Y., Huet, S. and Laurent, B. (2005). Testing convex hypotheses on the mean of a Gaussian vector. Application to testing qualitative hypotheses on a regression function. *Ann. Statist.* **33** 214–257. MR2157802

[6] Barlow, R. E., Bartholomew, D. J., Bremmer, J. M. and Brunk, H. D. (1972). *Statistical Inference under Order Restrictions*. Wiley, New York.

[7] Bartholomew, D. J. (1959). A test of homogeneity for ordered alternatives. *Biometrika* **46** 36–48. MR0104312

[8] Bellec, P. C. (2018). Sharp oracle inequalities for least squares estimators in shape restricted regression. *Ann. Statist.* **46** 745–780. MR3782383

[9] Bellec, P. C. and Tsybakov, A. B. (2015). Sharp oracle bounds for monotone and convex regression through aggregation. *Journal of Machine Learning Research* **16** 1879–1892. MR3417801

[10] Birgé, L. and Massart, P. (1993). Rates of convergence for minimum contrast estimators. *Probab. Theory Related Fields* **97** 113–150. MR1240719

[11] Bronshtein, E. M. (1976). ϵ-entropy of convex sets and functions. *Siberian Math. J.* **17** 393–398. MR0415155

[12] Brunk, H. D. (1955). Maximum Likelihood estimates of monotone parameters. *Ann. Math. Statist.* **26** 607–616. MR0073894
[13] Cator, E. (2011). Adaptivity and optimality of the monotone least-squares estimator. *Bernoulli* **17** 714–735. MR2787612
[14] Chatterjee, S. (2014). A new perspective on least squares under convex constraint. *Ann. Statist.* **42** 2340–2381. MR3269982
[15] Chatterjee, S. (2016). An improved global risk bound in concave regression. *Electron. J. Stat.* **10** 1608–1629. MR3522655
[16] Chatterjee, S., Guntuboyina, A. and Sen, B. (2015). On risk bounds in isotonic and other shape restricted regression problems. *Ann. Statist.* **43** 1774–1800. MR3357878
[17] Chatterjee, S., Guntuboyina, A. and Sen, B. (2018). On matrix estimation under monotonicity constraints. *Bernoulli* **24** 1072-1100. MR3706788
[18] Chen, Y. and Samworth, R. J. (2016). Generalized additive and index models with shape constraints. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 729–754. MR3534348
[19] Cule, M. and Samworth, R. (2010). Theoretical properties of the log-concave maximum likelihood estimator of a multidimensional density. *Electron. J. Stat.* **4** 254–270. MR2645484
[20] Cule, M. L., Samworth, R. J. and Stewart, M. I. (2010). Maximum likelihood estimation of a multi-dimensional log-concave density (with discussion). *J. Roy. Statist. Soc. Ser. B* **72** 545–600. MR2758237
[21] Donoho, D. L. (1990). Gelfand n–widths and the method of least squares. Technical report, Dept. Statistics, Univ. California, Berkeley.
[22] Dümbgen, L., Freitag–Wolf, S. and Jongbloed, G. (2006). Estimating a unimodal distribution from interval–censored data. *J. Amer. Statist. Assoc.* **101** 1094–1106. MR2324149
[23] Dümbgen, L., Rufibach, K. and Schuhmacher, D. (2014). Maximum–likelihood estimation of a log-concave density based on censored data. *Electron. J. Stat.* **8** 1405–1437. MR3263127
[24] Dümbgen, L., Samworth, R. and Schuhmacher, D. (2011). Approximation by log-concave distributions, with applications to regression. *Ann. Statist.* **39** 702–730. MR2816336
[25] Dümbgen, L., Wellner, J. A. and Wolff, M. (2016). A law of the iterated logarithm for Grenander's estimator. *Stochastic Process. Appl.* **126** 3854–3864. MR3565482
[26] Durot, C. (2007). On the L_p–error of monotonicity constrained estimators. *Ann. Statist.* **35** 1080–1104. MR2341699
[27] Durot, C. (2008). Monotone nonparametric regression with random design. *Math. Methods Statist.* **17** 327–341. MR2483461
[28] Dykstra, R. L. (1983). An algorithm for restricted least squares regression. *J. Amer. Statist. Assoc.* **78** 837–842. MR0727568
[29] Fraser, D. A. S. and Massam, H. (1989). A mixed primal-dual bases algorithm for regression under inequality constraints. Application to concave regression. *Scand. J. Statist.* **16** 65–74. MR1003969

[30] Gao, F. and Wellner, J. A. (2007). Entropy estimate for high-dimensional monotonic functions. *J. Multivariate Anal.* **98** 1751–1764. MR2392431
[31] Gao, F. and Wellner, J. A. (2017). Entropy of convex functions on $\mathbb{R}^d$. *Constr. Approx.* **46** 565–592. MR3735701
[32] Ghosal, S., Sen, A. and van der Vaart, A. W. (2000). Testing monotonicity of regression. *Ann. Statist.* **28** 1054–1082. MR1810919
[33] Grant, M. and Boyd, S. (2014). CVX: Matlab Software for Disciplined Convex Programming, version 2.1. http://cvxr.com/cvx.
[34] Grenander, U. (1957). On the theory of mortality measurement: II. *Skand. Aktuarietidskr.* **39** 125–153. MR0093415
[35] Groeneboom, P., Jongbloed, G. and Wellner, J. A. (2001). Estimation of a convex function: Characterization and asymptotic theory. *Ann. Statist.* **29** 1653–1698. MR1891742
[36] Groeneboom, P. and Jongbloed, G. (2010). Generalized continuous isotonic regression. *Statist. Probab. Lett.* **80** 248–253. MR2575453
[37] Guntuboyina, A. and Sen, B. (2013). Global risk bounds and adaptation in univariate convex regression. *Probab. Theory Related Fields* 1–33. MR3405621
[38] Hall, P. and Heckman, N. E. (2000). Testing for monotonicity of a regression mean by calibrating for linear functions. *Ann. Statist.* **28** 20–39. MR1762902
[39] Hall, P. and Huang, L. S. (2001). Nonparametric kernel regression subject to monotonicity constraints. *Ann. Statist.* **29** 624–647. MR1865334
[40] Han, Q. and Wellner, J. A. (2016). Multivariate convex regression: Global risk bounds and adaptation. arXiv:1601.06844.
[41] Han, Q., Wang, T., Chatterjee, S. and Samworth, R. J. (2017). Isotonic regression in general dimensions. arxiv:1708.09468. MR3988762
[42] Hannah, L. A. and Dunson, D. B. (2013). Multivariate convex regression with adaptive partitioning. *J. Mach. Learn. Res.* **14** 3261–3294. MR3144462
[43] Hanson, D. L., Pledger, G. and Wright, F. T. (1973). On consistency in monotonic regression. *Ann. Statist.* **1** 401–421. MR0353540
[44] Hanson, D. L. and Pledger, G. (1976). Consistency in concave regression. *Ann. Statist.* **4** 1038–1050. MR0426273
[45] Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*. Chapman & Hall/CRC, London. MR1082147
[46] Hildreth, C. (1954). Point estimates of ordinates of concave functions. *J. Amer. Statist. Assoc.* **49** 598–619. MR0065093
[47] Hull, J. C. (2006). *Options, Futures, and Other Derivatives*. Prentice Hall, New Jersey.
[48] Kim, A. K. H. and Samworth, R. J. (2016). Global rates of convergene in log-concave density estimation. *Ann. Statist.* **44** 2756–2779. MR3576560
[49] Koenker, R. and Mizera, I. (2010). Quasi–concave density estimation. *Ann. Statist.* **38** 2998–3027. MR2722462
[50] Kuosmanen, T. (2008). Representation theorem for convex nonparametric

least squares. *Econometrics J.* **11** 308–325.
[51] Kyng, R., Rao, A. and Sachdeva, S. (2015). Fast, provable algorithms for isotonic regression in all l_p–norms. *In Advances in Neural Information Processing Systems* 2719–2727.
[52] Lee, C., Johnson, A. L., Moreno-Centeno, E. and Kuosmanen, T. (2013). A More Efficient Algorithm for Convex Nonparametric Least Squares. *European J. Oper. Res.* **227** 391–400. MR3244897
[53] Lim, E. and Glynn, P. W. (2012). Consistency of Multidimensional Convex Regression. *Oper. Res.* **60** 196–208. MR2911667
[54] Luenberger, D. G. (1968). *Optimization by Vector Space Methods*. John Wiley & Sons, Inc., New York. MR0238472
[55] Makowski, G. G. (1977). Consistency of an estimator of doubly nondecreasing regression functions. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **39** 263–268. MR0652723
[56] Mammen, E. and Yu, K. (2007). Additive isotone regression. *IMS Lecture Notes–Monograph Series Asymptotics: Particles, Processes, and Inverse Problems* **55** 179–195. MR2459939
[57] Meyer, M. C. (2013). Semi–parametric additive constrained regression. *J. Nonparametr. Stat.* **25** 715–730. MR3174293
[58] Meyer, M. and Woodroofe, M. (2000). On the degrees of freedom in shape–restricted regression. *Ann. Statist.* **28** 1083–1104. MR1810920
[59] Nemirovski, A. M., Polyak, B. T. and Tsybakov, A. B. (1985). Rate of convergence of nonparametric estimators of maximum–likelihood type. *Problems of Information Transmission* **21** 258–272. MR0820705
[60] Patilea, V. (2001). Convex models, MLE and misspecification. *Ann. Statist.* **29** 94–123. MR1833960
[61] Rao, B. L. S. P. (1969). Estimation of a unimodel density. *Sankhyā Ser. A* **31** 23–36. MR0267677
[62] Robertson, T. and Wright, F. T. (1975). Consistency in generalized isotonic regression. *Ann. Statist.* **3** 350–362. MR0365871
[63] Robertson, T., Wright, F. T. and Dykstra, R. L. (1988). *Order Restricted Statistical Inference (Wiley Series in Probability and Mathematical Statistics)*. John Wiley & Sons, Chichester. MR0961262
[64] Schuhmacher, D. and Dümbgen, L. (2010). Consistency of multivariate log-concave density estimators. *Statist. Probab. Lett.* **80** 376–380. MR2593576
[65] Seijo, E. and Sen, B. (2011). Nonparametric least squares estimation of a multivariate convex regression function. *Ann. Statist.* **39** 1633–1657. MR2850215
[66] Sen, P. K. and Silvapulle, M. J. (2002). An appraisal of some aspects of statistical inference under inequality constraints. *J. Statist. Plann. Inference* **107** 3–43. MR1927753
[67] Seregin, A. and Wellner, J. A. (2010). Nonparametric estimation of multivariate convex-transformed densities. *Ann. Statist.* **38** 3751–3781. MR2766867
[68] Shapiro, A. (1988). Towards a unified theory of inequality constrained

testing in multivariate analysis. *Int. Stat. Rev.* **56** 49–62. MR0963140
[69] Stout, Q. F. (2015). Isotonic regression for multiple independent variables. *Algorithmica* **71** 450–470. MR3331888
[70] van de Geer, S. A. (1987). A new approach to least-squares estimation, with applications. *Ann. Statist.* **15** 587–602. MR0888427
[71] van de Geer, S. A. (1990). Estimating a regression function. *Ann. Statist.* **18** 907–924. MR1056343
[72] van de Geer, S. A. (1993). Hellinger–consistency of certain nonparametric maximum likelihood estimators. *Ann. Statist.* **21** 14–44. MR1212164
[73] van de Geer, S. (2000). *Applications of Empirical Process Theory*, Cambridge series in statistical and probabilistic mathematics ed. Cambridge University Press, Cambridge. MR1739079
[74] van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes with Applications to Statistics*. Springer, New York. MR1385671
[75] Wright, F. T. (1981). The asymptotic behavior of monotone regression estimates. *Ann. Statist.* **9** 443–448. MR0606630
[76] Yatchew, A. J. (1992). Nonparametric regression tests based on least squares. *Econom. Theory* **8** 435–451. MR1202324
[77] Zhang, C. H. (2002). Risk bounds in isotonic regression. *Ann. Statist.* **30** 528–555. MR1902898