Consistency of Multi–dimensional Convex Regression

Eunji Lim  
Industrial Engineering, University of Miami, Coral Gables, FL 33124, lim@miami.edu  
Peter W. Glynn  
Management Science and Engineering, Stanford University, Stanford, CA 94305, glynn@stanford.edu

Convex regression is concerned with computing the best fit of a convex function to a data set of \( n \) observations in which the independent variable is (possibly) multi–dimensional. Such regression problems arise in operations research, economics, and other disciplines in which imposing a convexity constraint on the regression function is natural. This paper studies a least squares estimator that is computable as the solution of a quadratic program and establishes that it converges almost surely to the “true” function as \( n \to \infty \) under modest technical assumptions. In addition to this multi–dimensional consistency result, we identify the behavior of the estimator when the model is mis–specified (so that the “true” function is non–convex) and extend the consistency result to settings in which the function must be both convex and non–decreasing (as is needed for consumer preference utility functions).

**Key words:**

**History:**

1. Introduction

This paper is concerned with the problem of convex regression in multiple dimensions. In particular, suppose that we observe \((X_1, Y_1), \ldots, (X_n, Y_n)\) and presume that

\[
Y_i = f_*(X_i) + \nu_i
\]

for \( i \geq 1 \), where \( f_* : \mathbb{R}^d \to \mathbb{R} \) is convex, with the “noise” \( \nu_i \) satisfying \( E\nu_i = 0 \) and \( E\nu_i^2 < \infty \). Our goal is to estimate the function \( f_* \) from the observed data; the estimator \( \hat{g}_n(\cdot) \) is the minimizer of the sum of squares

\[
\sum_{i=1}^{n} (Y_i - g(X_i))^2
\]  

(1.1)

over functions \( g : [0,1]^d \to \mathbb{R} \) that are convex. Our main result (Theorem 1) is that under modest technical assumptions, the estimator \( \hat{g}_n(\cdot) \) converges almost surely (a.s.) to \( f_* \) as \( n \to \infty \). In one
dimension, the theory of convex regression is well established; see Hanson and Pledger (1976) for consistency and Mammen (1991) and Groeneboom et al (2001) for rates of convergence results. Our main contribution is therefore the extension of consistency from one dimension to multiple dimensions. An important element in our results is that the number of samples per observation point is fixed at one. The almost sure consistency at a given $x$ (at which there may be no sample points at all, or at most one) is enforced by the shape constraint, in conjunction with information extracted from the entire sample. In fact, an interesting property of the estimator $\hat{g}_n$ is that an observation $(X_i, Y_i)$ with $X_i$ far from $x$ can have a significant impact on $\hat{g}_n(x)$, so that $\hat{g}_n(x)$ typically depends on the global structure of the sample (rather than only on “local samples” obtained near $x$, as with (for example) kernel density estimators; see, for example, Rosenblatt (1956)).

The development of a multidimensional nonparametric theory for convex regression is both natural from a statistical viewpoint and well motivated by operations research and economic applications. In the operations research setting, a number of different models give rise to associated performance measure expectations that are provably convex in the underlying model parameters; see Chapter 3 of Chen and Yao (2001) for a discussion of such results in the queueing network context. When the performance measure expectation is computed via (Monte Carlo) simulation, one is then led to a convex regression problem. On the other hand, in the economics context, it is often presumed that utility function preferences are described by a concave function $u$ (or, equivalently, that $-u$ is convex). When utility information is solicited empirically from consumers, a convex regression statistical formulation is a natural model to consider as a mechanism for estimating $u$; see, for example, Meyer and Pratt (1968). Such convexity constraints can also arise in estimation of supply and demand functions.

Because one of our interests is in applying these methods to the analysis of simulation output from complex stochastic models (as in Lim and Glynn (2006)), it will often be the case that convex regression will be applied in settings in which one has no a priori guarantee that the function $f_*$ is convex. Our theory therefore permits the possibility of model mis-specification, and studies the behavior of our estimator when $f_*$ is non-convex. In this case, our estimator $\hat{g}_n$ converges to
the convex \( g^* \) that is closest to \( f^* \) in a certain \( L^2 \) space; see Section 2 for details. Our proof of convergence makes critical use of the fact that the class of convex functions is itself a convex subset of the space of functions.

This paper can also be viewed as a contribution to the larger literature on regression in the presence of “shape constraints” on the regression function. The earliest (and most extensively studied) such problem is that of isotone regression (in which the regression function \( f^* \) is presumed to be “monotone”); see, for example, Brunk (1958), Barlow and Brunk (1972), and Wright (1979). To our knowledge, this is the first such paper in the shape–constrained literature to consider the possibility of model mis–specification and to note the important role that the space \( L^2(X) \) (to be introduced in Section 2) plays in such problems.

Our paper is organized as follows. Section 2 introduces the mathematical framework for our analysis, and precisely states the main theorem (Theorem 1) in this paper. The proof of this result is provided in Section 3, while Section 4 discusses a couple of extensions of our ideas. In particular, we discuss the extension of our convergence result to multi–dimensional convex regression problems in which the domain of the convex function is a convex subset of \( \mathbb{R}^d \), and to problems in which the function \( f^* \) is assumed to be both convex and non–decreasing. This latter extension is particularly relevant to estimation of customer preference functions, given that preferences are generally assumed to be non–decreasing in the underlying “baskets” of goods.

2. The Main Result

The framework for our analysis presumes that we observe \( n \) pairs \((X_1,Y_1), \ldots, (X_n,Y_n)\), in which \( X_i \) is a continuous \( \mathbb{R}^d \)–valued “independent variable observation” and \( Y_i \) is the corresponding real–valued “dependent variable observation”. We allow for the possibility of using a positive continuous “weight function” \( w \), so that (1.1) in general is replaced by the “sum of squares”

\[
\varphi_n(g) \triangleq \frac{1}{n} \sum_{i=1}^{n} w(X_i) (Y_i - g(X_i))^2.
\]

Given the above “goodness of fit” criterion, our goal is to estimate \( f^* \) by minimizing \( \varphi_n \) over \( \mathcal{C} = \{ g : \mathbb{R}^d \to \mathbb{R} \text{ such that } g \text{ is convex} \} \). Since \( \mathcal{C} \) is infinite–dimensional, this minimization may appear
to be computationally intractable (without introducing a further finite-dimensional approximation).

However, it turns out that this minimization can be formulated as a finite-dimensional quadratic program (QP).

**Proposition 1.** Consider the quadratic program (in the decision variables \((g_1, \xi_1), \ldots, (g_n, \xi_n)\))

\[
\min \frac{1}{n} \sum_{i=1}^{n} w(X_i) (Y_i - g_i)^2 \tag{2.1}
\]

\[s/t \ g_j \geq g_i + \xi_i^T (X_j - X_i), \quad 1 \leq i, j \leq n.\]

For \(n \geq d+1\), the QP (2.1) has a minimizer \((\tilde{g}_1, \tilde{\xi}_1), \ldots, (\tilde{g}_n, \tilde{\xi}_n)\) and the minimizing values \(\tilde{g}_1, \ldots, \tilde{g}_n\) are unique. Furthermore, any minimizer \(\tilde{g}_n\) of \(\varphi_n\) over \(C\) satisfies \(\tilde{g}_n(X_i) = \tilde{g}_i\) for \(1 \leq i \leq n\).

**Proof.** Clearly, the constraints of (2.1) describe a set that is convex in \(g = (g_1, \ldots, g_n)^T\). In addition, we claim that the set of such \(g\)'s is non-empty (choose \(g_j = 0 = \xi_j\) for \(1 \leq j \leq n\)) and closed. To verify the closedness property, suppose that \(((g^k_i : 1 \leq i \leq n) : k \geq 1)\) is a sequence (in \(k\)) of elements in the set for which \(g^k_i \to g^\infty_i\) as \(k \to \infty\) (for \(1 \leq i \leq n\)). We need to show that the limit \((g^\infty_i : 1 \leq i \leq n)\) possesses a finite-valued set of subgradients satisfying the constraints of (2.1). Fix \(i\) and let \((\xi^k_i : k \geq 1)\) be the sequence of corresponding subgradients at \(X_i\). If \((\xi^k_i : k \geq 1)\) is a bounded sequence, we can extract a convergent subsequence and define a valid \(\xi^\infty_i\) as its limit point. On the other hand, if \(\|\xi^k_i\|\) converges to infinity, define \(\tilde{\xi}_i^k = \xi^k_i/\|\xi^k_i\|\). Since the \(\tilde{\xi}_i^k\)'s are bounded, we can again extract a convergent subsequence, and define \(\tilde{\xi}^\infty_i\) as its limit. If \((\tilde{\xi}^\infty_i)^T (X_j - X_i)\) is non-zero for some \(j\), then it is evident that \((g^k_j - g_j^k : k \geq 1)\) cannot be bounded if \(\|\xi^k_i\| \to \infty\) as \(k \to \infty\), establishing a contradiction. Finally, if \(\tilde{\xi}^\infty_i\) is orthogonal to \(X_j - X_i\) for \(1 \leq j \leq d+1\), we again arrive at a contradiction since \((X_i - X_l : l \neq i, 1 \leq l \leq d+1)\) is a.s. a collection of linearly independent random vectors spanning \(R^d\) (under our assumption that the \(X_i\)'s are continuous random vectors). The last part of the proposition follows from p. 337 of Boyd and Vandenberghe (2004). □

While Proposition 1 asserts that \((\tilde{g}_1, \ldots, \tilde{g}_n)\) is unique, there are many convex functions \(g\) satisfying \(g(X_i) = \tilde{g}_i\) for \(1 \leq i \leq n\). To define our estimator \(\hat{g}_n(x)\) at \(x \neq X_i, 1 \leq i \leq n\), we set

\[
\hat{g}_n(x) = \sup\{g(x) : g \in C, g(X_i) = \tilde{g}_i, 1 \leq i \leq n\}. \tag{2.2}
\]
The function $\hat{g}_n(\cdot)$ is guaranteed to be convex; see, for example, p. 75 of Avriel (1976). Furthermore, $\hat{g}_n(\cdot)$ is finite–valued on $\text{conv}(X_1,\ldots,X_n)$ (where $\text{conv}(A) \triangleq$ convex hull of $A$, for $A \subset \mathbb{R}^n$), and infinite elsewhere.

In principle, the maximization that defines (2.2) again appears to be infinite–dimensional (over $C$). However, as with Proposition 1, the next result makes clear that $\hat{g}_n(x)$ can be computed as the solution of a finite–dimensional problem (this time, a linear program (LP)).

**Proposition 2.** For each $x \in \mathbb{R}^d$, $\hat{g}_n(x)$ can be computed as the optimal value $\hat{y}$ of the LP (in the decision variables $y, \xi_1,\ldots,\xi_n, \tilde{\xi}$)

$$\max y \quad \text{s/t} \quad \begin{align*}
\hat{g}_j &\geq \hat{g}_i + \xi_i^T(X_j - X_i), \quad 1 \leq i,j \leq n \\
y &\geq \hat{g}_i + \xi_i^T(x - X_i), \quad 1 \leq i \leq n \\
\hat{g}_j &\geq y + \tilde{\xi}_i^T(X_j - x), \quad 1 \leq j \leq n.
\end{align*}$$

**Proof.** The constraints of (2.3) describe piece–wise affine convex functions taking values $g_1,\ldots,g_n,y$ at $X_1,\ldots,X_n,x$ (with corresponding subgradients $\xi_1,\ldots,\xi_n, \tilde{\xi}$). Since this is a sub-class of the functions appearing on the right–hand side of (2.2), $\hat{y} \leq \hat{g}_n(x)$. On the other hand, every finite–valued convex function agreeing with the $\hat{g}_i$’s at the $X_i$’s and taking on value $y$ at $x$ must possess a corresponding set of subgradients satisfying (2.3), so that $\hat{y} = \hat{g}_n(x)$ (We refer to p. 34 of Rockafellar (1974) for the fact that the set of subgradients must be non–empty at $X_1,\ldots,X_n,x$). □

The convex function $\hat{g}_n(\cdot)$ is our estimator for $f_*(\cdot)$. In order to analyze this estimator, we shall impose some probabilistic assumptions on the $(X_i,Y_i)$’s. In particular, we require that:

A1. $X,X_1,X_2,\ldots$ is a sequence of independent and identically distributed (iid) $\mathbb{R}^d$–valued random vectors having a common continuous positive density $k(\cdot)$.

A2. For $i \geq 1$, $Y_i = f_*(X_i) + \nu_i$, where the $\nu_i$’s satisfy

$$P(\nu_i \in dy_i,1 \leq i \leq n|X_1,X_2,\ldots) = \prod_{i=1}^n F(dy_i|X_i)$$
for some family $(F(\cdot|x): x \in \mathbb{R}^d)$ of cumulative distribution functions.

A3. $Ew(X_1)(Y_1^2 + \|X_1\|^2 + 1) < \infty$, thereby implying that

$$\sigma^2(X_1) \triangleq \int_{\mathbb{R}} y^2 F(dy|X_1) < \infty \quad \text{a.s.}$$

A4. For each $x \in \mathbb{R}^d$,

$$\int_{\mathbb{R}} y F(dy|x) = 0.$$ Let $L^2(X) = \{g: \mathbb{R}^d \to \mathbb{R} \text{ such that } Ew(X)g^2(X) < \infty\}$. We equip $L^2(X)$ with the inner product

$$<g_1, g_2> = Ew(X)g_1(X)g_2(X)$$

and associated norm $\|g\|_2 = \sqrt{<g, g>}$ for $g_1, g_2, g \in L^2(X)$. In view of A3,

$$\infty > Ew(X_1)E[Y_1^2|X_1] \geq Ew(X_1) (E[Y_1|X_1]^2) = Ew(X_1)f_\alpha^2(X_1)$$

so $f_\alpha \in L^2(X)$. We allow for the possibility that the model is mis-specified, so that $f_\alpha$ is not in $\mathcal{C}$. To characterize the limiting behavior of $\hat{g}_n$ in the presence of model mis-specification, we note that

$$\mathcal{C}^2 = \{g \in \mathcal{C}: g \in L^2(X)\}$$

is a convex cone (i.e. $\mathcal{C}^2$ is convex and $g \in \mathcal{C}^2$ implies that $\alpha g \in \mathcal{C}^2$ for $\alpha \geq 0$).

**PROPOSITION 3.** $\mathcal{C}^2 \subset L^2(X)$ is a closed convex cone.

**Proof.** The only non-trivial issue is the verification that $\mathcal{C}^2$ is closed in $L^2(X)$. Suppose that $\|g_n - g_\infty\|_2 \to 0$ as $n \to \infty$, where $g_n \in \mathcal{C}^2$ for $n \geq 1$. Since $L^2(X)$ is complete, we need only show that $g_\infty \in \mathcal{C}$.

It is a standard fact that $g_\infty$ can be defined via the pointwise limit

$$g_\infty(x) = \lim_{n \to \infty} \sup g_n(x).$$

We shall now establish that $g_\infty$ is a convex function that is finite-valued everywhere. To prove convexity, we show that the epigraph of $g_\infty$, namely

$$P(g_\infty) \triangleq \{(x, \alpha) \in \mathbb{R}^d \times \mathbb{R}: g_\infty(x) \leq \alpha\}$$
is a convex subset of $\mathbb{R}^{d+1}$. If \((x_1, \alpha_1) \) and \((x_2, \alpha_2) \in P(g_\infty)\), it follows that for each $\epsilon > 0$, there exists $n_0$ such that for $n \geq n_0$,

$$g_n(x_1) \leq \alpha_1 + \epsilon$$

and

$$g_n(x_2) \leq \alpha_2 + \epsilon$$

for $n \geq n_0$. But $g_n$ is convex so its epigraph is convex, and hence

$$g_n(px_1 + (1-p)x_2) \leq p(\alpha_1 + \epsilon) + (1-p)(\alpha_2 + \epsilon)$$

(2.4)

for $0 \leq p \leq 1$. Consequently,

$$g_\infty(px_1 + (1-p)x_2) = \limsup_{n \to \infty} g_n(px_1 + (1-p)x_2) \leq p\alpha_1 + (1-p)\alpha_2 + \epsilon.$$ 

Since $\epsilon > 0$ was arbitrary, we conclude that $(px_1 + (1-p)x_2, p\alpha_1 + (1-p)\alpha_2) \in P(g_\infty)$, proving the convexity of $P(g_\infty)$.

The effective domain of $g_\infty$, namely $ED(g_\infty) \triangleq \{x : g_\infty(x) < \infty\}$ is necessarily convex. On the other hand, because $g_\infty \in L^2(X)$ and $w$ and $k$ are positive, $R^d - ED(g_\infty)$ must be a set of (Lebesgue) measure 0. Consequently, $ED(g_\infty) = R^d$. Suppose now that $g_\infty(x_0) = -\infty$ for some $x_0 \in R^d$. Then, $g_\infty(x) = -\infty$ for all $x \in R^d$; see Theorem 4.16 of Avriel (1976), for example. Since $g_\infty$ is finite–valued a.e., this contradiction proves that $g_\infty$ is finite–valued everywhere. □

Since $C^2$ is a closed convex set contained in $L^2(X)$, there exists a unique function $g_* \in C^2$ which is the minimizer of

$$\min_{g \in C^2} \|f_* - g\|_2.$$ 

Furthermore, $g_* \in C^2$ is characterized either by the pair of relations

$$< f_* - g_*, g_*> = 0 \quad \text{and} \quad < f_* - g_*, g> \leq 0$$

(2.5)

for $g \in C^2$, or (equivalently) via the relation

$$< f_* - g_*, g - g_* \geq 0$$

(2.6)
for $g \in C^2$; see Brunk (1965) for details. Note that (2.5) guarantees that $\|g_*\|^2_2 \leq \|f_*\|^2_2 \leq EY^2_1 w(X_1)$.

We are now ready to state the main result in this paper.

**Theorem 1.** Assume A1–A4. Then, for each $c \geq 0$,

$$\sup_{\|x\| \leq c} |\hat{g}_n(x) - g_*(x)| \to 0 \quad \text{a.s.}$$

as $n \to \infty$, where $\|x\| = \max(|x_i| : 1 \leq i \leq d)$ for $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$.

This theorem is the key consistency result in this paper concerning the estimator $\hat{g}_n$. Note that if $f_* \in C$ (so that the model is not mis-specified), the result asserts that $\hat{g}_n$ converges uniformly to $f_*$ on compact sets. Uniform convergence on $\mathbb{R}^d$ typically fails, because the subgradients close to the boundary of $\text{conv}(X_1, \ldots, X_n)$ may be badly behaved.

The next section of this paper is concerned with supplying the details of the proof of Theorem 1.

**3. Proof of the Main Result**

Our proof of Theorem 1 can be broken down into a number of key steps.

**Step 1** Exploit the fact that $\hat{g}_n$ is the minimizer of the sum of squares: Observe that since $\hat{g}_n$ is a minimizer of $\varphi_n$ over $C$, it follows that $\varphi_n(\hat{g}_n) \leq \varphi_n(g_*)$. Furthermore,

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

$$= \frac{1}{n} \sum_{i=1}^n (Y_i - g_*(X_i))^2 w(X_i)$$

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

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

The inequality $\varphi_n(\hat{g}_n) \leq \varphi_n(g_*)$ therefore yields the bound

$$\frac{1}{n} \sum_{i=1}^n (\hat{g}_n(X_i) - g_*(X_i))^2 w(X_i) \leq \frac{2}{n} \sum_{i=1}^n (Y_i - g_*(X_i)) (\hat{g}_n(X_i) - g_*(X_i)) w(X_i). \quad (3.1)$$

The right–hand side of (3.1) is essentially a sample version of the inner product (2.6); the difficulty in exploiting (2.6) directly is that $\hat{g}_n$ is not a fixed (deterministic) function. (If $\hat{g}_n$ were a fixed
deterministic function in \( L^2(X) \), the strong law could be directly applied and (2.6) immediately invokes thereby verifying (3.13) and allowing us to skip Steps 2 through 8.)

**Step 2** Obtain a bound on the “empirical \( L^2 \) norm” of \( \hat{g}_n \): Applying the Cauchy–Schwarz inequality path–by–path to the right–hand side of (3.1), we find that

\[
\frac{1}{n} \sum_{i=1}^{n} (\hat{g}_n(X_i) - g_*(X_i))^2 w(X_i) \leq 2 \sqrt{\frac{1}{n} \sum_{i=1}^{n} (Y_i - g_*(X_i))^2 w(X_i) \cdot \frac{1}{n} \sum_{i=1}^{n} (\hat{g}_n(X_i) - g_*(X_i))^2 w(X_i)}
\]

so that

\[
\frac{1}{n} \sum_{i=1}^{n} (\hat{g}_n(X_i) - g_*(X_i))^2 w(X_i) \leq \frac{4}{n} \sum_{i=1}^{n} (Y_i - g_*(X_i))^2 w(X_i).
\] (3.2)

Since \((a + b)^2 \leq 2a^2 + 2b^2\) for \(a, b \in \mathbb{R}\), we may conclude that

\[
\frac{1}{n} \sum_{i=1}^{n} \hat{g}_n(X_i) w(X_i) \leq \frac{2}{n} \sum_{i=1}^{n} (\hat{g}_n(X_i) - g_*(X_i))^2 w(X_i) + 2 \sum_{i=1}^{n} g_*(X_i)^2 w(X_i) \leq \frac{8}{n} \sum_{i=1}^{n} (Y_i - g_*(X_i))^2 w(X_i) + 2 \sum_{i=1}^{n} g_*(X_i)^2 w(X_i).
\]

Because the \((X_i, Y_i)\)'s are iid, the strong law of large numbers ensures that

\[
\frac{1}{n} \sum_{i=1}^{n} \hat{g}_n^2(X_i) w(X_i) \leq 9E (Y_1 - g_*(X_1))^2 w(X_1) + 3Eg_*(X_1)^2 w(X_1) \doteq \beta < \infty \quad \text{(3.3)}
\]
a.s. for \(n\) sufficiently large.

**Step 3** Show that the contribution to the empirical inner product appearing on the right–hand side of (3.1) from \(X_i\)'s lying outside an appropriately chosen compact set can be made arbitrarily small. Specifically, we note that a path–by–path application of the Cauchy–Schwarz inequality establishes that

\[
\frac{1}{n} \sum_{i=1}^{n} (Y_i - g_*(X_i)) (\hat{g}_n(X_i) - g_*(X_i)) w(X_i) I(\|X_i\| > c) \leq \sqrt{\frac{1}{n} \sum_{i=1}^{n} (Y_i - g_*(X_i))^2 w(X_i) I(\|X_i\| > c)} \cdot \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\hat{g}_n(X_i) - g_*(X_i))^2 w(X_i)}
\]
\[
\leq \sqrt{2E (Y_i - g_*(X_i))^2 w(X_i) I (\|X_i\| > c)} \cdot \left( \frac{4}{n} \sum_{i=1}^{n} (Y_i - g_*(X_i))^2 w(X_i) \right)^{1/2} \\
\leq \sqrt{2E (Y_i - g_*(X_i))^2 w(X_i) I (\|X_i\| > c)} \cdot \left( \sqrt{5E (Y_i - g_*(X_i))^2 w(X_i)} + 1 \right) \tag{3.4}
\]

for \( n \) sufficiently large (where we used (3.2) for the second last inequality). Since \( EY_i^2 w(X_i) < \infty \) and \( g_* \in L^2(X) \), the right-hand side of (3.4) can be made smaller than any \( \epsilon > 0 \) by choosing \( c \) sufficiently large.

**Step 4** Use the empirical \( L^2 \) norm bound and convexity to bound the maximum of \( |\hat{g}_n| \) on compact sets. Specifically, we will prove the following result.

**Proposition 4.** Assume A1–A4. Then, for each \( c \geq 0 \), there exists a deterministic constant \( \tilde{\beta}(c) < \infty \) such that

\[
\limsup_{n \to \infty} \sup_{\|x\| \leq c} |\hat{g}_n(x)| \leq \tilde{\beta}(c) \quad a.s.
\]

**Proof.** Let \( e = (1, 1, \ldots, 1)^T \), \( e_i \) be the \( i \)'th unit vector \( (1 \leq i \leq d) \), and \( e_0 = (0, 0, \ldots, 0)^T \). We will prove that for each \( c \geq 0 \),

\[
\limsup_{n \to \infty} \sup_{x \in \mathcal{H}} |\hat{g}_n(x)|
\]

is a.s. bounded by a deterministic constant, where \( \mathcal{H} = \{ x : \|x - (c/2d)e\| \leq c/8d \} \); the general case then follows via a translation of \( \mathbb{R}^d \) to \( \mathbb{R}^d + (c/2d)e \) (and re-scaling \( c \) suitably). The proof proceeds by using the empirical \( L^2 \) bound to bound the maximum of \( |\hat{g}_n| \) over a neighborhood of the points \( \{ce_0, \ldots, ce_d, (c/2d)e\} \); the convexity of \( \hat{g}_n \) then guarantees that \( |\hat{g}_n| \) is bounded over \( \mathcal{H} \).

Let \( B_i = \{ x \in \mathbb{R}^d_i : \|x - ce_i\| \leq \tau c/2d \} \) for \( 1 \leq i \leq d \) and \( B_{d+1} = \{ x \in \mathbb{R}^d : \|x - (c/2d)e\| \leq \tau c/8d \} \), where \( 0 < \tau \leq 1/2 \) will be determined later. For each \( B_i \), let \( \gamma_i = (1/2)P(X \in B_i) \) and set \( \gamma = \min \{ \gamma_i : 0 \leq i \leq d \} \). Then, for \( 0 \leq i \leq d+1 \) and \( r > 0 \),

\[
\frac{1}{n} \sum_{j=1}^{n} I (X_j \in B_i, |\hat{g}_n(X_j)| \leq r) \\
\geq \frac{1}{n} \sum_{j=1}^{n} I (X_j \in B_i) - \frac{1}{n} \sum_{j=1}^{n} I (X_j \in B_i, |\hat{g}_n(X_j)| > r) \\
\geq \frac{1}{n} \sum_{j=1}^{n} I (X_j \in B_i) - \frac{1}{n} \sum_{j=1}^{n} I (X_j \in B_i, |\hat{g}_n(X_j)| > r) w(X_j)/ \inf_{x \in B_i} w(x)
\]
\[ \sum_{j=1}^{n} I(X_j \in B_i) - \frac{1}{w(x)} \sum_{j=1}^{n} I(X_j \in B_i, |\hat{g}_n(X_j)| > r) w(X_j). \quad (3.5) \]

But Markov’s inequality and (3.3) imply that
\[ \frac{1}{n} \sum_{j=1}^{n} I(|\hat{g}_n(X_j)| > r) w(X_j) \leq r^{-2} \frac{1}{n} \sum_{j=1}^{n} \hat{g}_n(X_j)^2 w(X_j) \leq r^{-2} \beta \quad (3.6) \]
for \( n \) sufficiently large. Choose \( r_0 \) so large that \( r_0^{-2} \beta \leq \gamma / 2 \). It follows from (3.5) and (3.6) that
\[ \frac{1}{n} \sum_{j=1}^{n} I(X_j \in B_i, |\hat{g}_n(X_j)| \leq r_0) \geq \gamma / 2 \quad (3.7) \]
for \( n \) sufficiently large. For each such \( n \), there therefore exists \( X_{I(i)} \in B_i \) with \( 1 \leq I(i) \leq n \) and \( |\hat{g}_n(X_{I(i)})| \leq r_0 \). For each \( x \) in the convex hull of \( \{X_{I(0)}, \ldots, X_{I(d)}\} \), \( x = \sum_{i=0}^{d} p_i X_{I(i)} \) for some convex combination \( p_0, p_1, \ldots, p_d \), so that the convexity of \( \hat{g}_n \) yields
\[ \hat{g}_n(x) \leq \sum_{i=0}^{d} p_i \hat{g}_n(X_{I(i)}) \leq r_0. \quad (3.8) \]

We now show that the convex hull contains \( \tilde{H} = \{x : \|x - (c/2)d e\| \leq c/4d\} \). We need to prove that there exist non–negative \( p_0, p_1, \ldots, p_d \) summing to one such that
\[ \sum_{i=0}^{d} p_i X_{I(i)} = x \]
for \( x \in \tilde{H} \). This is equivalent to requiring a non–negative solution \( p_1, \ldots, p_d \) (summing to less than or equal to one) to the linear system
\[ \sum_{i=1}^{d} p_i \left( c e_i + \tilde{X}_{I(i)} - X_{I(0)} \right) = x - X_{I(0)}, \quad (3.9) \]
where \( \tilde{X}_{I(i)} = X_{I(i)} - c e_i \) for \( 1 \leq i \leq d \). The linear system (3.9) can be re–expressed as
\[ (I + F/c) p = \frac{1}{c} \left( x - X_{I(0)} \right), \]
where \( F = (F_{ij} : 1 \leq i, j \leq d) \) is a square \( d \times d \) matrix in which the \( i \)'th column is \( \tilde{X}_{I(i)} - X_{I(0)} \). Set
\[ |||F||| \equiv \max_{1 \leq i \leq d} \sum_{j=1}^{d} |F_{ij}| \]
and note that $||c^{-1}F|| \leq \tau$. Hence, $(I + c^{-1}F)^{-1}$ exists and $||(I + c^{-1}F)^{-1}|| \leq (1 - ||c^{-1}F||)^{-1} \leq (1 - \tau)^{-1} \leq 2$. So,

$$p - x/c = -c^{-1}X_{I(0)} + c^{-1}F(I - c^{-1}F)^{-1}(x - X_{I(0)})c^{-1}$$

and hence

$$\|p - x/c\| \leq \tau/2d + 2\tau \leq 3\tau$$

if $\|x\| \leq c$. It follows that if $x_i/c \geq 3\tau$ ($1 \leq i \leq d$) and $c^{-1}\sum_{i=1}^{d}x_i \leq 1 - 3\tau d$, the $p_i$'s solving (3.9) are non-negative and sum to less than or equal to one, so that $x = (x_1, \ldots, x_d)^T$ is in the convex hull of the $X_{I(i)}$'s. If we choose $\tau$ so that $\tau \leq (12d)^{-1}$, it is easily seen that each $x \in \tilde{H}$ satisfies these linear equalities, so that $\tilde{H}$ is contained in the convex hull. Relation (3.6) then implies that

$$\sup_{x \in \tilde{H}} \hat{g}_n(x) \leq r_0$$

(3.10)

for $n$ sufficiently large.

To obtain a lower bound on $\hat{g}_n(\cdot)$ over $\mathcal{H}$, we use $X_{I(d+1)}$. Suppose that $x$ is such that

$$X_{I(d+1)} = 1/2x + 1/2y$$

(3.11)

for $y \in \tilde{H}$. Then, the convexity of $\tilde{g}_n(\cdot)$ establishes the inequality

$$\hat{g}_n(X_{I(d+1)}) \leq 1/2\hat{g}_n(x) + 1/2\hat{g}_n(y),$$

so that for $n$ large

$$\hat{g}_n(x) \geq 2\hat{g}_n(X_{I(d+1)}) - \hat{g}_n(y) \geq 2\hat{g}_n(X_{I(d+1)}) - r_0 \geq -3r_0.$$ 

So the lower bound ensues if we can write $x$ as in (3.11). Observe that

$$y - (c/2d)e = 2(X_{I(d+1)} - (c/2d)e) - (x - (c/2d)e)$$
so

\[ \| y - (c/2d)e \| \leq \tau c/4d + \| x - (c/2d)e \| \leq c/8d + \| x - (c/2d)e \|. \]

Thus, \( y \in \tilde{H} \) (i.e. \( \| y - (c/2d)e \| \leq c/4d \)) if \( x \in H \) (i.e. \( \| x - (c/2d)e \| \leq c/8d \)), proving that

\[ \inf_{x \in H} \hat{g}_n(x) \geq -3r_0. \]  

(3.12)

Combining (3.10) and (3.12) proves the result. \( \square \)

**Step 5** Set \( \mathcal{H}_c = \{ x : \| x \| \leq c \} \). Observe that the a.s. bound on \( |\hat{g}_n(\cdot)| \) (uniformly in \( n \) over \( \mathcal{H}_{c(1+\delta)} \)) implies that \( \hat{g}_n \) is Lipschitz over \( \mathcal{H}_c \) uniformly in \( n \) a.s. In particular, 

\[ |\hat{g}_n(x) - \hat{g}_n(y)| \leq (2/(c\delta))\tilde{\beta}(c(1+\delta)) \| x - y \| \]

for \( x, y \in \mathcal{H}_c \) for \( n \) sufficiently large; see, for example, Van der Vaart and Wellner (2000), p. 165, and Roberts and Varberg (1974).

**Step 6** Let

\[ \mathcal{C}_c = \{ h : \mathcal{H}_c \to R \text{ such that } h \text{ is convex on } \mathcal{H}_c, \]

\[ |h(x)| \leq \tilde{\beta}(c), \ |h(x) - h(y)| \leq 2/(c\delta)\tilde{\beta}(c(1+\delta)) \| x - y \| \text{ for } x, y \in \mathcal{H}_c \} \]

and note that Steps 4 and 5 guarantee that for each \( c \geq 0 \), \( \hat{g}_n \in \mathcal{C}_c \) for \( n \) sufficiently large a.s. Furthermore, \( \mathcal{C}_c \) is compact in the uniform metric \( d_c \) given by

\[ g_c(h_1, h_2) = \sup_{x \in \mathcal{H}_c} |h_1(x) - h_2(x)|. \]

It follows that for each \( \epsilon > 0 \), there exists a finite collection of convex functions \( h_1, h_2, \ldots, h_m \) such that \( \bigcup_{i=1}^m \{ h \in \mathcal{C}_c : d_c(h_i, h) < \epsilon \} \supseteq \mathcal{C}_c \) (i.e. \( h_1, h_2, \ldots, h_m \) is an \( \epsilon \)-net for \( \mathcal{C}_c \)). In fact, \( \log m \) is of order \( \epsilon^{-d/2} \) as \( \epsilon \downarrow 0 \) (for fixed \( c \)); see Theorem 6 of Bronshtein (1976).

**Step 7** Fix \( \epsilon > 0 \) and choose \( c \) so large that the right-hand side of (3.4) is less than \( \epsilon \). For \( \delta > 0 \), let \( h_1, h_2, \ldots, h_m \) be an \( \epsilon \)-net for \( \mathcal{C}_{c(1+\delta)} \). For each \( h_i \), we show that we can extend \( h_i \) to a finite-valued convex function \( \tilde{h}_i \) on \( R^d \) that agrees with \( h_i \) on \( \mathcal{C}_c \) and lies in \( L^2(X) \).
To prove this, observe first that because $h_i$ is Lipschitz on $H_{c(1+\delta)}$, it is continuous at each $x \in H_c$ (including limits taken from outside $H_c$). Theorem 11 of Rockafellar (1974) therefore applies, so that the set of subgradients for $h_i$ is non-empty at each $x \in H_c$. But each subgradient $\partial h_i(x)$ satisfies
\[ h_i(x + re_j) \geq h_i(x) + r \partial h_i(x)^T e_j \]
for $1 \leq j \leq d$ and $r$ in some neighborhood of 0. Letting $\eta(c)$ be a bound on the Lipschitz constant for functions $h \in C_c(1+\delta)$, we find that $|h_i(x + re_j) - h_i(x)| \leq \eta(c)|r|$ for $r$ sufficiently small, so that
\[ \max_{1 \leq j \leq d} |(\partial h_i(x))_j| \leq \eta(c). \]
Set
\[ \tilde{h}_i(y) = \sup_{x \in H_c} h_i(x) + \partial h_i(x)^T (y - x). \]
Note that $\tilde{h}_i(x) = h_i(x)$ for $x \in H_c$, is convex, and satisfies
\[ |\tilde{h}_i(y) - \tilde{h}_i(x)| \leq \eta(c)\|x - y\| \]
for all $x, y \in R^d$, so that $\tilde{h}_i$ is globally Lipschitz. Since $\tilde{h}_i$ grows linearly at $\infty$ and is bounded on compact sets, $\tilde{h}_i \in L^2(X)$ by virtue of A3.

**Step 8** We now use the empirical inner product inequality (3.1), and the fact that $\hat{g}_n$ can be uniformly approximated to precision $\epsilon$ within $H_c$, to conclude that
\[ \frac{1}{n} \sum_{i=1}^{n} (\hat{g}_n(X_i) - g_*(X_i))^2 w(X_i) \to 0 \quad (3.13) \]
a.s. as $n \to \infty$.

To fill in the details, fix $\epsilon > 0$, choose $c$ as in Step 7, and select $\tilde{h}_1, \tilde{h}_2, \ldots, \tilde{h}_m$ as suggested in Step 7. Then, for each $j \in \{1, \ldots, m\}$,
\[
\frac{1}{n} \sum_{i=1}^{n} (Y_i - g_*(X_i)) \left( \hat{g}_n(X_i) - g_*(X_i) \right) w(X_i) I(\|X_i\| \leq c) \\
\leq \frac{1}{n} \sum_{i=1}^{n} (Y_i - g_*(X_i)) \left( \hat{g}_n(X_i) - \tilde{h}_j(X_i) \right) w(X_i) I(\|X_i\| \leq c)
\]
As a consequence,

\[ \frac{1}{n} \sum_{i=1}^{n} (Y_i - g_*(X_i)) \left( \bar{h}_j(X_i) - g_*(X_i) \right) w(x_i) + \frac{1}{n} \sum_{i=1}^{n} |Y_i - g_*(X_i)| |\bar{h}_j(X_i) - g_*(X_i)| w(x_i) I (\|X_i\| > c). \]

\[ \leq \frac{1}{n} \sum_{i=1}^{n} |Y_i - g_*(X_i)| \cdot \sup_{x \in H_c} \left| \tilde{g}_n(x) - \tilde{h}_j(x) \right| w(x_i) I (\|X_i\| \leq c) + \max_{1 \leq k \leq m} \frac{1}{n} \sum_{i=1}^{n} (Y_i - g_*(X_i)) \left( \bar{h}_k(X_i) - g_*(X_i) \right) w(x_i) \]

\[ + \max_{1 \leq k \leq m} \frac{1}{n} \sum_{i=1}^{n} (Y_i - g_*(X_i))^2 w(x_i) I (\|X_i\| > c) \cdot \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left( \bar{h}_k(X_i) - g_*(X_i) \right)^2 w(x_i)}. \]

As a consequence,

\[ \frac{1}{n} \sum_{i=1}^{n} (Y_i - g_*(X_i)) (\tilde{g}_n(X_i) - g_*(X_i)) w(x_i) I (\|X_i\| \leq c) \leq \min \sup_{1 \leq k \leq m_{\epsilon}} \left| \tilde{g}_n(x) - \tilde{h}_j(x) \right| \cdot \frac{1}{n} \sum_{i=1}^{n} |Y_i - g_*(X_i)| w(x_i) I (\|X_i\| \leq c) + \max_{1 \leq k \leq m} \frac{1}{n} \sum_{i=1}^{n} (Y_i - g_*(X_i)) \left( \tilde{h}_k(X_i) - g_0(X_i) \right) w(x_i) \]

\[ + \max_{1 \leq k \leq m} \frac{1}{n} \sum_{i=1}^{n} (Y_i - g_*(X_i))^2 w(x_i) I (\|X_i\| > c) \cdot \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left( \tilde{h}_k(X_i) - g_*(X_i) \right)^2 w(x_i)}. \]

\[ \leq \epsilon \cdot \frac{1}{n} \sum_{i=1}^{n} |Y_i - g_*(X_i)| w(x_i) \]

\[ + \max_{1 \leq k \leq m} \frac{1}{n} \sum_{i=1}^{n} (Y_i - g_*(X_i)) \left( \tilde{h}_k(X_i) - g_*(X_i) \right) w(x_i) \]

\[ + \max_{1 \leq k \leq m} \frac{1}{n} \sum_{i=1}^{n} (Y_i - g_*(X_i))^2 w(x_i) I (\|X_i\| > c) \cdot \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left( \tilde{h}_k(X_i) - g_*(X_i) \right)^2 w(x_i)}, \]

where we used the fact that \( \tilde{h}_1, \ldots, \tilde{h}_m \) is an \( \epsilon \)-net for \( C_c \) in the last inequality. Because \( \tilde{h}_1, \ldots, \tilde{h}_m \) are in \( L^2(X) \), the strong law of large numbers for iid sequences guarantees that

\[ \limsup_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} (Y_i - g_*(X_i)) (\tilde{g}_n(X_i) - g_*(X_i)) w(x_i) I (\|X_i\| \leq c) \]

\[ \leq \epsilon E |Y_1 - g_*(X_1)| w(x_1) \]

\[ + \max_{1 \leq k \leq m} E (Y_1 - g_*(X_1)) \left( \tilde{h}_k(X_1) - g_*(X_1) \right) w(x_1) \]

\[ + \max_{1 \leq k \leq m} \sqrt{E (Y_1 - g_*(X_1))^2 w(x_1)} I (\|X_1\| > c) \cdot \sqrt{E \left( \tilde{h}_k(X_1) - g_*(X_1) \right)^2 w(x_1)}. \]
\[ \leq c\sqrt{E(Y_1 - g_*(X_1))^2 w(X_1)} \cdot \sqrt{Ew(X_1)} + \max_{1 \leq k \leq m} \epsilon < f_* - g_*, \tilde{h}_k - g_* > \\
+ \epsilon \max_{1 \leq k \leq m} \sqrt{E\left(\tilde{h}_k(X_1) - g_*(X_1)\right)^2 w(X_1)} \]

a.s. Because \( \tilde{h}_k \in C^2 \) for \( 1 \leq k \leq n \), (2.5) ensures that \( < f_* - g_*, \tilde{h}_k - g_* > \leq 0 \) for \( 1 \leq k \leq m \). In view of Step 3, we therefore conclude that

\[ \limsup_{n \to \infty} \frac{1}{n} \left( Y_i - g_*(X_i) \right) (\tilde{g}_n(X_i) - g_*(X_i)) w(X_i) \]

\[ \leq c\epsilon \sqrt{E(Y_1 - g_*(X_1))^2 w(X_1)} \cdot \sqrt{Ew(X_1)} \]

\[ + \epsilon \max_{1 \leq k \leq m} \sqrt{E\left(\tilde{h}_k(X_1) - g_*(X_1)\right)^2 w(X_1)} \]

a.s. Since \( \epsilon > 0 \) was arbitrary, (3.1) implies that

\[ \limsup_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} (\tilde{g}_n(X_i) - g_*(X_i))^2 w(X_i) \leq 0 \text{ a.s.,} \]

yielding (3.13).

**Step 9** We now use (3.13) and the fact that \( \tilde{g}_n \) is Lipschitz (uniformly in \( n \)) over each compact set \( \mathcal{H}_c \) to conclude that \( \tilde{g}_n \) converges to \( g_* \) uniformly a.s. over \( \mathcal{H}_c \).

Fix \( \epsilon > 0 \). Since \( \mathcal{H}_c \) is compact, we can find a finite collection of sets \( \Lambda_1, \ldots, \Lambda_l \) covering \( \mathcal{H}_c \), each having diameter less than \( \epsilon \) (i.e. \( \sup \{ \|x - y\| : x, y \in \Lambda_j \} \leq \epsilon \)). According to Step 5, we can find a (uniform in \( n \)) Lipschitz constant, call it \( \lambda(c) \), for both \( \tilde{g}_n \) and \( g_* \) over \( \mathcal{H}_c \). For each \( x \in \Lambda_j \) and \( X_i \in \Lambda_j \),

\[ |\tilde{g}_n(x) - g_*(x)| \leq |\tilde{g}_n(x) - \tilde{g}_n(X_i)| + |\tilde{g}_n(X_i) - g_*(X_i)| + |g_*(X_i) - g_*(x)| \]

\[ \leq \lambda(c) \epsilon + |\tilde{g}_n(X_i) - g_*(X_i)| + \lambda(c) \epsilon, \]

so

\[ \sup_{x \in B_j} |\tilde{g}_n(x) - g_n(x)| \]

\[ \leq 2\lambda(c) \epsilon + \sum_{i=1}^{n} |\tilde{g}_n(X_i) - g_*(X_i)| I(X_i \in \Lambda_j) \]

\[ \leq 2\lambda(c) \epsilon + \frac{1}{n} \sum_{i=1}^{n} |\tilde{g}_n(X_i) - g_*(X_i)| w(X_i) \cdot \sum_{i=1}^{n} \frac{n}{I(X_i \in \Lambda_j)} \]
\[ \leq 2\lambda(c)\epsilon + \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\hat{g}_n(X_i) - g_*(X_i))^2 w(X_i)} \cdot \frac{1}{n} \sum_{i=1}^{n} w(X_i) \cdot \frac{n}{\sum_{i=1}^{n} I(X_i \in \Lambda_j)}. \]

Taking advantage of (3.13), we conclude that

\[ \limsup_{n \to \infty} \sup_{x \in \Lambda_j} |\hat{g}_n(x) - g_*(x)| \leq 2\lambda(c)\epsilon \quad \text{a.s.} \]

Since \(\epsilon\) was arbitrary and there are only finitely many \(\Lambda_j\)'s, we conclude that

\[ \sup_{x \in H_c} |\hat{g}_n(x) - g_*(x)| \to 0 \quad \text{a.s.} \]

as \(n \to \infty\), proving Theorem 1.

4. Extensions

We consider here two extensions of the convex regression problem discussed in Sections 2 and 3.

Extension 1: The domain of the convex function to be estimated need not be \(\mathbb{R}^d\).

In some applications (e.g. that of estimating a utility function), the natural domain is not \(\mathbb{R}^d\), but some convex subset of \(\mathbb{R}^d\) (e.g. \(\mathbb{R}^d_+\)). In such settings, the proof of Theorem 1 carries over (with the natural proviso that \(w\) and \(k\) continue to be strictly positive and continuous on the interior of the domain), with the conclusion that the estimator \(\hat{g}_n\) converges uniformly a.s. to \(g_*\) on compact subsets that are contained in the interior of the domain. In general, uniform convergence fails at the boundaries of the domain, because there are very few observations that lie close to the boundaries, and the convex function may go to infinity at the boundary (even when the domain is compact).

If the domain is compact, say \([0, 1]^d\), and one desires uniform convergence on \([0, 1]^d\) (rather than on compact subsets of the interior), one will typically need to require that a significant fraction of the sampling occur at all the extreme points of the domain \([0, 1]^d\). For example, if the sampling distribution for \(X\) is a strictly positive mixture of a continuous positive probability density function on \([0, 1]^d\) and the \(2^d\) point masses on the extreme points of \([0, 1]^d\), then the \(L^2\) bound of Step 2 will easily imply that \(\hat{g}_n(\cdot)\) is bounded (uniformly in \(n\)) at the \(2^d\) extreme points of \([0, 1]^d\) and at \((1/2)\epsilon\), so that \(\hat{g}_n\) is Lipschitz (uniformly in \(n\)) over \([0, 1]^d\) (by virtue of Roberts and Varberg (1974)).
The rest of the argument follows as in Section 3 (except that it is easier, since (for example) the functions $h_j$ arising in Step 6 need not be “extended to infinity”).

**Extension 2:** Adding the requirement that the function to be estimated be both convex and non-decreasing

In the context of estimating utility functions and supply/demand functions, it is natural to impose the requirement that the function not only be convex/concave, but that it is also non-decreasing. In particular, we say that a function $g : R^d \rightarrow R$ is non-decreasing if $g(x) \leq g(y)$ whenever $x \leq y$ (so that $x_i \leq y_i$ for $1 \leq i \leq d$). We now adapt the definition of the cone of functions $\mathcal{C}$ to $\mathcal{C} = \{ g : R^d \rightarrow R \text{ such that } g \text{ is convex and non-decreasing} \}$, and assume $A1$–$A4$. We can now analogously define $L^2(X)$ and $C^2$ as in Section 2.

Given a function $f_*$, our estimator $\hat{g}_n$ is again obtained by solving a QP:

$$\min \frac{1}{n} \sum_{i=1}^{n} w(X_i)(Y_i - g_i)^2$$

s.t. $g_j \geq g_i + \xi^T_i(X_j - X_i), \quad 1 \leq i,j \leq n$

$$\xi_i \geq 0, \quad 1 \leq i \leq n.$$ 

As in Section 2, this defines $\hat{g}_n(\cdot)$ at the $X_i$’s. To define $\hat{g}_n(\cdot)$ globally, we again define $\hat{g}_n(\cdot)$ via (2.2) (but with our modified definition of $\mathcal{C}$); it is easily seen that $\hat{g}_n$ is convex and non-decreasing on $\text{conv}(X_1, \ldots, X_n)$ ( since the supremum of convex non-decreasing functions is necessarily convex and non-decreasing). In addition, for $x \neq X_i, 1 \leq i \leq n$, $\hat{g}_n(x)$ can be evaluated by solving an LP:

$$\max \ y$$

s.t. $\hat{g}_j \geq \hat{g}_i + \xi^T_i(X_j - X_i), \quad 1 \leq i,j \leq n$

$$y \geq \hat{g}_i + \xi^T_i(x - X_i), \quad 1 \leq i \leq n$$

$$\hat{g}_j \geq y + \tilde{\xi}^T(X_j - x), \quad 1 \leq j \leq n$$

$$\tilde{\xi} \geq 0, \xi_i \geq 0, \quad 1 \leq i \leq n.$$ 

Because $\mathcal{C}$ is a subcone of the cone of convex functions used in Section 3, Steps 1 through 5 of Section 3 follow as in that section. For Step 6, modify $\mathcal{C}_c$ so that the requirement that $h$ is...
non-decreasing is added to the definition provided in Section 3. Clearly, $C_c$ is a bounded and equi-
continuous family of functions. Furthermore, any pointwise limit of functions lying in $C_c$ must be
convex, non-decreasing, and satisfy the Lipschitz and boundedness constraints, so that $C_c$ is closed.
It follows that $C_c$ is compact in the metric $d_c$, and hence is totally bounded; see Copson (1972),
Chapter 6, for example. Thus, for each $\epsilon > 0$, there exists a finite $\epsilon$-net $h_1, h_2, \ldots, h_m$ of elements
in $C_c$ for which their corresponding $\epsilon$-balls cover $C_c$.

Step 7 again requires extending $h_i$ from $\mathcal{H}_c$ to a convex non-decreasing function $\tilde{h}_i$ on $\mathbb{R}^d$.
Because $h_i$ is non-decreasing, it follows that for $x \in \mathcal{H}_c$, $h_i(x - re_j) \leq h_i(x)$ for $r$ small and positive.
But
\[
h_i(x - re_j) \geq h_i(x) - r\partial h_i(x)^T e_j,
\]
from which we conclude that $\partial h_i(x)_j \geq 0$, and hence $\partial h_i(x) \geq 0$. Thus, for each $x \in \mathcal{H}_c$, $h_i(x) +
\partial h_i(x)^T (y - x)$ is a non-decreasing convex function, so that
\[
\tilde{h}_i(y) \triangleq \sup_{x \in \mathcal{H}_c} h_i(x) + \partial h_i(x)^T (y - x)
\]
is convex, non-decreasing, and agrees with $h_i$ on $\mathcal{H}_c$. Furthermore, as in Section 3, $\tilde{h}_i$ is globally
Lipschitz, and therefore lies in $L^2(\mathbb{X})$ as a consequence of A3. Steps 8 and 9 follow the same lines
as in Section 3, so that we therefore obtain the following theorem.

**Theorem 2.** Assume A1–A4. Then for each $c \geq 0$,
\[
\sup_{\|x\| \leq c} |\hat{g}_n(x) - g_*(x)| \to 0 \quad a.s.
\]
as $n \to \infty$, where $g_*$ is characterized either by (2.5) or (2.6) and $C^2 \subset L^2(\mathbb{X})$ is the class of convex
non-decreasing functions.

**Acknowledgments**

The research of the second author was partially supported by a grant from the King Abdullah University of
Science and Technology.
References


