OPERATIONS RESEARCH
Vol. 60, No. 1, January–February 2012, pp. 196–208
ISSN 0030-364X (print) | ISSN 1526-5463 (online)

http://dx.doi.org/10.1287/opre.1110.1007

Consistency of Multidimensional Convex Regression

Eunji Lim
Department of Industrial Engineering, University of Miami, Coral Gables, Florida 33124,
lim@miami.edu

Peter W. Glynn
Department of Management Science and Engineering, Stanford University, Stanford, California 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) multidimensional. 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 multidimensional consistency result, we identify the behavior of the estimator when the model is misspecified (so that the "true" function is nonconvex), and we extend the consistency result to settings in which the function must be both convex and nondecreasing (as is needed for consumer preference utility functions).

Subject classifications: nonparametric regression; multidimensional convex functions; asymptotic properties; consistency.
Area of review: Simulation.
History: Received January 2010; revisions received September 2010, December 2010, March 2011; accepted April 2011.

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

$$Y_i = f_*(X_i) + \nu_i \tag{1}$$

for $i \geqslant 1$, where $f_*: \mathbb{R}^d \to \mathbb{R}$ is convex, with the "noise" ν_i satisfying $\mathbb{E}[\nu_i] = 0$ and $\mathbb{E}[\nu_i^2] < \infty$. 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

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

over functions $g: \mathbb{R}^d \to \mathbb{R}$ that are convex. Our main result (Theorem 1) is that under modest technical assumptions, the least-squares estimator $\hat{g}_n(\cdot)$ converges almost surely (a.s.) to f_* as $n \to \infty$. In one dimension, the theory of least-squares based 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 for such least-squares estimators from one dimension to multiple dimensions. An important element in our results is that the number of samples per observation point is fixed at 1. 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 such least-squares estimators 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; see, for example, Allon et al. (2007), as well as prior contributions of Varian (1984, 1985), Hall and Huang (2001), and Yatchew and Bos (1997).

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 misspecification, and studies the behavior of our estimator when f_* is nonconvex. 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 §2 for details.

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). A major advantage of shape-based function estimation relative to other nonparametric function estimation methods is that no smoothing parameters need be specified (e.g., the kernel density bandwidth). The main prior contribution to this area is Allon et al. (2007), in which a convex function is fit nonparametrically to statistical data using the method of nonparametric maximum likelihood (in contrast to the least-squares estimator that is the subject of this paper). The key differences between our contribution and Allon et al. (2007) are as follows:

1. In our formulation, the error appears additively (as in (1)), whereas Allon et al. (2007) assume that the error enters multiplicatively; see §5 for additional discussion. Our additive error formulation follows the prevailing literature for shape-based regression used historically.

2. The nonparametric maximum-likelihood estimator developed in Allon et al. (2007) assumes that the density of the logarithm of the noise is known, is symmetric about its mean, is log-concave, and is common across all values of the independent variable x. As shown in §5, these assumptions are necessary, in the sense that their estimator can converge to an incorrect limit when their assumptions are violated. By contrast, our development allows both the underlying convex function and the distribution of the noise to be specified nonparametrically. Furthermore, we permit the distribution of the noise to vary arbitrarily as a function of the independent variable. Note that these assumptions, together with the presumption of additive error, fit the simulation setting of interest well.

3. The convergence proof of Allon et al. (2007) presumes a known bound on the global Lipschitz constant of the underlying convex function, which then enters the definition of the estimator as well. In some sense, such a presumption is tantamout to specifying a smoothing constant for the estimator (which, of course, shape-based regression is intended to avoid). On the other hand, our Theorem 1 assumes no a priori bound on the global Lipschitz constant of the underlying convex function. In fact, given that the typical convex function is globally non-Lipschitz, our estimator does not even assume that the function is globally Lipschitz. It should be noted that our proof would be much simpler if we imposed such a condition.

4. The presumption of convexity is one that imposes infinitely many constraints on the behavior of a function. In view of this, model misspecification is an important issue both theoretically and practically. Our contribution (Theorem 2) analyzes the large-sample behavior of the estimator when the underlying function is nonconvex (under additional shape restrictions). This continues a tradition of recent literature on shape-based regression, specifically work by Dümbgen et al. (2011) and Cule and Samworth (2010) in which model misspecification in the setting of nonparametric maximum-likelihood estimation of log-concave densities has been analyzed. In this density estimation context, the misspecified estimator converges to that log-concave density that is closest to the underlying density in the sense of Kullback-Liebler divergence (which has no obvious generalization to the nondensity convex functions that are considered by us).

5. Our proof of consistency, in contrast to previous arguments applied within the shape-constrained literature, adopts a Hilbert space perspective (and uses L^2 arguments rather than arguments based on the topology of uniform convergence on compact sets). In particular, our proof takes advantage of the presence of an inner product, as well as the fact that the space of appropriately integrable convex functions forms a closed convex subset within our chosen Hilbert space, to obtain a geometric characterization of the limit point to which our estimator converges; this characterization plays a key role in our model misspecification theory. We believe that our use of Hilbert space ideas may be valuable in other shape-based contexts as well. In fact, we show in §5 how our ideas can easily be extended to least-squares estimation in the presence of a shape constraint involving functions that are both convex and nondecreasing.

In the process of preparing a second revision of this paper, one of the referees brought to our attention an independent contribution to the problem of convex regression, authored by Seijo and Sen (2011). The theory developed there, although using somewhat different proof techniques, does not consider the problem of model misspecification.

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 §3, whereas §4 discusses a couple of extensions of our ideas. In particular, we discuss the extension of our convergence result to multidimensional 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 nondecreasing. This latter extension is particularly relevant

to estimation of customer preference functions, given that preferences are generally assumed to be nondecreasing in the underlying "baskets" of goods. Finally, §5 contrasts our estimator against the Allon et al. (2007) estimator in greater detail, and discusses computational experiences with these estimators.

2. The Main Results

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) 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 φ_n over $\mathscr{C} = \{g\colon \mathbb{R}^d \to \mathbb{R} \text{ such that } g \text{ is convex}\}$. Because $\mathscr{C}$ 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); see Kuosmanen (2008) for details, and also Boyd and Vandenberghe (2004, p. 338). A related finite-dimensional reduction can be found in Allon et al. (2007).

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

$$\begin{aligned} \min\ & \frac{1}{n}\sum_{i=1}^{n} w(X_i)(Y_i - g_i)^2 \qquad (3)\\ \text{s/t}\ & g_j \geqslant g_i + \xi_i^T(X_j - X_i), \quad 1 \leqslant i, j \leqslant n. \end{aligned}$$

For $n \geqslant d + 1$, *the QP* (3) *has a minimizer* $(\hat{g}_1, \hat{\xi}_1), \ldots, (\hat{g}_n, \hat{\xi}_n)$ *and the minimizing values* $\hat{g}_1, \ldots, \hat{g}_n$ *are unique. Furthermore, any minimizer* $\tilde{g}_n$ *of* φ_n *over* $\mathscr{C}$ *satisfies* $\tilde{g}_n(X_i) = \hat{g}_i$ *for* $1 \leqslant i \leqslant n$.

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

$$\hat{g}_n(x) = \sup\{g(x)\colon g \in \mathscr{C},\ g(X_i) = \hat{g}_i,\ 1 \leqslant i \leqslant n\}. \qquad (4)$$

Note that $\hat{\xi}_1, \ldots, \hat{\xi}_n$ are then subgradients of the convex function $\hat{g}_n(\cdot)$ at the points $X_1, \ldots, X_n$. Furthermore, $\hat{g}_n(\cdot)$ is finite valued on $\mathrm{conv}(X_1, \ldots, X_n)$ (where $\mathrm{conv}(A) \triangleq$ convex hull of A, for $A \subset \mathbb{R}^n$), and infinite elsewhere.

In principle, the maximization that defines (4) again appears to be infinite dimensional (over $\mathscr{C}$). 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, $\xi_1, \ldots, \xi_n, \tilde{\xi}$)

$$\begin{aligned} \max\ & y \qquad (5)\\ \text{s/t}\ & \hat{g}_j \geqslant \hat{g}_i + \xi_i^T(X_j - X_i), \quad 1 \leqslant i, j \leqslant n\\ & y \geqslant \hat{g}_i + \xi_i^T(x - X_i), \quad 1 \leqslant i \leqslant n\\ & \hat{g}_j \geqslant y + \tilde{\xi}^T(X_j - x), \quad 1 \leqslant j \leqslant n. \end{aligned}$$

Proof. The constraints of (5) describe piecewise-affine convex functions taking values $g_1, \ldots, g_n, y$ at $X_1, \ldots, X_n, x$ (with corresponding subgradients $\xi_1, \ldots, \xi_n, \tilde{\xi}$). Because this is a subclass of the functions appearing on the right-hand side of (4), $\hat{y} \leqslant \hat{g}_n(x)$. On the other hand, every finite-valued convex function agreeing with the $\hat{g}_i$s at the X_is and taking on value y at x must possess a corresponding set of subgradients satisfying (5), 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 nonempty 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 the following:

A1. $X, X_1, X_2, \ldots$ is a sequence of independent and identically distributed (i.i.d.) $\mathbb{R}^d$-valued random vectors having a common distribution with support $\mathbb{R}^d$ (so that $P(X \in A) > 0$ for each open subset $A \subset \mathbb{R}^d$).

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

$$P(\nu_i \in dy_i, 1 \leqslant i \leqslant n \mid X_1, X_2, \ldots) = \prod_{i=1}^{n} F(dy_i \mid X_i)$$

for some family $(F(\cdot \mid x)\colon x \in \mathbb{R}^d)$ of cumulative distribution functions.

A3. $\mathbb{E}[w(X_1)(Y_1^2 + \|X_1\|^2 + 1)] < \infty$, thereby implying that

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

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

$$\int_{\mathbb{R}} yF(dy \mid x) = 0.$$

To understand the large-sample behavior of our estimator $\hat{g}_n$, note that for $g \in \mathscr{C}$

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

Because $\mathbb{E}[\nu_i(f_*(X_i) - g(X_i))w(X_i)] = 0$, the second term converges to 0 a.s. as $n \to \infty$. On the other hand, the first term converges to $\mathbb{E}[(f_*(X) - g(X))^2 w(X)]$ and the third term converges to $\mathbb{E}[\nu_1^2 w(X_1)]$ a.s. as $n \to \infty$. Because $\hat{g}_n$ minimizes $\sum_{i=1}^n (Y_i - g(X_i))^2 w(X_i)/n$ (which, in turn, converges to $\mathbb{E}[(f_*(X) - g(X))^2 w(X)] + \mathbb{E}[\nu_1^2 w(X_1)]$), we expect that $\hat{g}_n$ should converge to the minimizer of $\mathbb{E}[(f_*(X) - g(X))^2 w(X)]$ over $g \in \mathscr{C}$.

To make this precise, we use the Hilbert space $L^2 = \{g\colon \mathbb{R}^d \to \mathbb{R} \text{ such that } \mathbb{E}[w(X)g^2(X)] < \infty\}$. We equip L^2 with the inner product

$$\langle g_1, g_2 \rangle = \mathbb{E}[w(X)g_1(X)g_2(X)]$$

and associated norm $\|g\|_2 = \sqrt{\langle g, g \rangle}$ for $g_1, g_2, g \in L^2$. In view of A3,

$$\begin{aligned}\infty > \mathbb{E}[w(X_1)]\mathbb{E}[Y_1^2 \mid X_1] &\geqslant \mathbb{E}[w(X_1)]\mathbb{E}[Y_1 \mid X_1]^2 \\ &= \mathbb{E}[w(X_1)f_*^2(X_1)]\end{aligned}$$

so $f_* \in L^2$. We are now ready to state our main convergence result (covering the case where the model has been correctly specified, so that $f_* \in \mathscr{C}$).

THEOREM 1. *Assume* A1–A4 *and that* $f_* \in \mathscr{C}$. *Then, for each* $c \geqslant 0$,

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

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

This theorem is the key consistency result in this paper concerning the estimator $\hat{g}_n$. Uniform convergence on $\mathbb{R}^d$ typically fails, because the subgradients close to the boundary of $\operatorname{conv}(X_1, \ldots, X_n)$ may be badly behaved.

We now turn to the question of what occurs when the model is misspecified, so that f_* is not in $\mathscr{C}$. To characterize the limiting behavior of $\hat{g}_n$ in the presence of model misspecification, we note that

$$\mathscr{C}^2 = \{g \in \mathscr{C}\colon g \in L^2\}$$

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

PROPOSITION 3. $\mathscr{C}^2 \subset L^2$ *is a closed convex cone.*

PROOF. The only nontrivial issue is the verification that $\mathscr{C}^2$ is closed in L^2. Suppose that $\|g_n - g_\infty\|_2 \to 0$ as $n \to \infty$, where $g_n \in \mathscr{C}^2$ for $n \geqslant 1$. Because L^2 is complete, we need only show that $g_\infty \in \mathscr{C}$.

It is a standard fact that g_∞ can be defined via the pointwise limit

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

We shall now establish that g_∞ is a convex function that is finite valued everywhere. To prove convexity, note that $\sup_{k \geqslant n} g_k$ is convex (see, for example, Theorem 4.13 of Avriel 1976). Because convexity is presumed under pointwise limits, g_∞ is therefore convex.

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

Because $\mathscr{C}^2$ is a closed convex set contained in L^2, there exists a unique function $g_* \in \mathscr{C}^2$ which is the minimizer of $\min_{g \in \mathscr{C}^2} \|f_* - g\|_2$.

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

$$\langle f_* - g_*, g_* \rangle = 0 \quad \text{and} \quad \langle f_* - g_*, g \rangle \leqslant 0 \tag{6}$$

for $g \in \mathscr{C}^2$, implying that

$$\langle f_* - g_*, g - g_* \rangle \leqslant 0 \tag{7}$$

for $g \in \mathscr{C}^2$; see Brunk (1965) for details. Note that (6) guarantees that $\|g_*\|_2^2 \leqslant \|f_*\|_2^2 \leqslant \mathbb{E}[Y_1^2 w(X_1)]$. Although we believe that $\hat{g}_n$, as defined above, converges a.s. to g_* under A1–A4, our current proof method for studying the impact of model misspecification requires a slight modification of the problem (and hence the estimator). In particular, we now assume that it is believed that f_* is nonnegative and is such that $f_* \leqslant k$, where $k \in L^2$ is a known function. In view of these additional shape constraints, we incorporate this information into our definition of the estimator $\hat{g}_n$. To define the modified estimator, we first solve the convex optimization problem

$$\begin{aligned}\min \quad & \frac{1}{n}\sum_{i=1}^n w(X_i)(Y_i - g_i)^2 \qquad (8)\\ \text{s/t} \quad & g_j \geqslant g_i + \xi_i^T(X_j - X_i), \quad 1 \leqslant i, j \leqslant n \\ & k(x) \geqslant g_i + \xi_i^T(x - X_i) \quad x \in \mathbb{R}^d \\ & g_i \geqslant 0, \quad 1 \leqslant i \leqslant n,\end{aligned}$$

having minimizing values $\hat{g}_1, \ldots, \hat{g}_n$. The estimator $\hat{g}_n$ is then defined globally (at a given $x \in \mathbb{R}^d$) via the maximization problem (in the decision variables $y, \tilde{\xi}, \xi_1, \xi_2, \ldots, \xi_n$)

$$\begin{aligned}\max \quad & y \qquad (9)\\ \text{s/t} \quad & \hat{g}_j \geqslant \hat{g}_i + \xi_i^T(X_j - X_i), \quad 1 \leqslant i, j \leqslant n \\ & y \geqslant \hat{g}_i + \xi_i^T(x - X_i), \quad 1 \leqslant i \leqslant n \\ & \hat{g}_j \geqslant y + \tilde{\xi}^T(X_j - x), \quad 1 \leqslant j \leqslant n \\ & k(z) \geqslant \hat{g}_i + \xi_i^T(z - X_i), \quad 1 \leqslant i \leqslant n,\ z \in \mathbb{R}^d \\ & k(z) \geqslant y + \tilde{\xi}^T(z - x), \quad z \in \mathbb{R}^d \\ & y \geqslant 0.\end{aligned}$$

Note that the estimator $\hat{g}_n$ is now guaranteed to be both nonnegative and satisfy $\hat{g}_n \leqslant k$. Let $\tilde{\mathscr{C}}^2 = \{g \in \mathscr{C}^2\colon\ 0 \leqslant g \leqslant k\}$, and observe that $\mathscr{C}^2$ is a closed convex subset of $\mathscr{C}^2$. We can now state our main result concerning the case of model misspecification.

THEOREM 2. *Assume* A1–A4, *and let* $\hat{g}_n$ *be defined via* (8) *and* (9). *If* $g_* \in \tilde{\mathscr{C}}^2$ *is the unique function for which* $\langle f_* - g_*, g - g_*\rangle \leqslant 0$ *for* $g \in \tilde{\mathscr{C}}^2$, *then for each* $c \geqslant 0$,

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

as $n \to \infty$.

Thus, if the model is misspecified, Theorem 2 assures us that our estimator converges uniformly a.s. on compact sets to the function g_* that is closest (in our L^2 distance) to the misspecified f_*.

We turn next to the proof section.

3. Proofs of Theorems 1 and 2

We first prove Theorem 1 and follow that with a proof of Theorem 2 (that leverages off the proof structure for Theorem 1). 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 because $\hat{g}_n$ is a minimizer of φ_n over $\mathscr{C}$, it follows that $\varphi_n(\hat{g}_n) \leqslant \varphi_n(g_*)$. Furthermore,

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

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

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^n (\hat{g}_n(X_i) - g_*(X_i))^2 w(X_i)\\ &\quad \leqslant \frac{2}{n}\sum_{i=1}^n (Y_i - g_*(X_i))(\hat{g}_n(X_i) - g_*(X_i))w(X_i). \qquad (10)\end{aligned}$$

The right-hand side of (10) is essentially a sample version of the inner product (10); the difficulty in exploiting (10) directly is that $\hat{g}_n$ is not a fixed (deterministic) function. (If $\hat{g}_n$ were a fixed deterministic function in L^2, the strong law could be directly applied and (10) immediately applies, thereby verifying (22) and allowing us to skip Steps 2 through 7.)

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 (10), we find that

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^n (\hat{g}_n(X_i) - g_*(X_i))^2 w(X_i)\\ &\leqslant 2\sqrt{\frac{1}{n}\sum_{i=1}^n (Y_i - g_*(X_i))^2 w(X_i) \cdot \frac{1}{n}\sum_{i=1}^n (\hat{g}_n(X_i) - g_*(X_i))^2 w(X_i)}\end{aligned}$$

so that

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^n (\hat{g}_n(X_i) - g_*(X_i))^2 w(X_i)\\ &\quad \leqslant \frac{4}{n}\sum_{i=1}^n (Y_i - g_*(X_i))^2 w(X_i). \qquad (11)\end{aligned}$$

Because $(a+b)^2 \leqslant 2a^2 + 2b^2$ for $a, b \in \mathbb{R}$, we may conclude that

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^n \hat{g}_n(X_i) w(X_i)\\ &\quad \leqslant \frac{2}{n}\sum_{i=1}^n (\hat{g}_n(X_i) - g_*(X_i))^2 w(X_i) + \frac{2}{n}\sum_{i=1}^n g_*(X_i)^2 w(X_i)\\ &\quad \leqslant \frac{8}{n}\sum_{i=1}^n (Y_i - g_*(X_i))^2 w(X_i) + \frac{2}{n}\sum_{i=1}^n g_*(X_i)^2 w(X_i).\end{aligned}$$

Because the (X_i, Y_i)'s are i.i.d.; the strong law of large numbers ensures that

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n \hat{g}_n^2(X_i) w(X_i) &\leqslant 9\mathbb{E}[(Y_1 - g_*(X_1))^2 w(X_1)]\\ &\quad + 3\mathbb{E}[g_*(X_1)^2 w(X_1)] \triangleq \beta < \infty \qquad (12)\end{aligned}$$

a.s. for n sufficiently large.

Step 3. Show that the contribution to the empirical inner product appearing on the right-hand side of (10) from X_is 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

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^n (Y_i - g_*(X_i))(\hat{g}_n(X_i) - g_*(X_i))w(X_i) I(\|X_i\| > c)\\ &\quad \leqslant \sqrt{\frac{1}{n}\sum_{i=1}^n (Y_i - g_*(X_i))^2 w(X_i) I(\|X_i\| > c)}\\ &\qquad \cdot \sqrt{\frac{1}{n}\sum_{i=1}^n (\hat{g}_n(X_i) - g_*(X_i))^2 w(X_i)}\\ &\quad \leqslant \sqrt{2\mathbb{E}[(Y_1 - g_*(X_1))^2 w(X_1) I(\|X_1\| > c)]}\\ &\qquad \cdot \sqrt{\frac{4}{n}\sum_{i=1}^n (Y_i - g_*(X_i))^2 w(X_i)}\\ &\quad \leqslant \sqrt{2\mathbb{E}[(Y_1 - g_*(X_1))^2 w(X_1) I(\|X_1\| > c)]}\\ &\qquad \cdot (\sqrt{5\mathbb{E}[(Y_1 - g_*(X_1))^2 w(X_1)]} + 1) \qquad (13)\end{aligned}$$

for n sufficiently large (where we used (11) for the second last inequality). Because $\mathbb{E}[Y_1^2 w(X_1)] < \infty$ and $g_* \in L^2$, the right-hand side of (13) 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* $c \geqslant 0$, *there exists a deterministic constant* $\tilde{\beta}(c) < \infty$ *such that*

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

PROOF. Let $e = (1, 1, \ldots, 1)^T$, e_i be the ith unit vector ($1 \leqslant i \leqslant d$), and $e_0 = (0, 0, \ldots, 0)^T$. We will prove that for each $c \geqslant 0$,

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

is a.s. bounded by a deterministic constant, where $\mathcal{H} = \{x\colon \|x - (c/2d)e\| \leqslant c/8d\}$; the general case then follows via a translation of $\mathbb{R}^d$ to $\mathbb{R}^d + (c/2d)e$ (and rescaling 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 $B_i = \{x \in \mathbb{R}_+^d\colon \|x - ce_i\| \leqslant \tau c/2d\}$ for $1 \leqslant i \leqslant d$ and $B_{d+1} = \{x \in \mathbb{R}_+^d\colon \|x - (c/2d)e\| \leqslant \tau c/8d\}$, where $0 < \tau \leqslant 1/2$ will be determined later. For each B_i, let $\gamma_i = (1/2)P(X \in B_i)$ and set $\gamma = \min\{\gamma_i\colon 0 \leqslant i \leqslant d\}$. Then, for $0 \leqslant i \leqslant d+1$ and $r > 0$,

$$\begin{aligned}
&\frac{1}{n}\sum_{j=1}^n I(X_j \in B_i,\ |\hat{g}_n(X_j)| \leqslant r)\\
&\quad\geqslant \frac{1}{n}\sum_{j=1}^n I(X_j \in B_i) - \frac{1}{n}\sum_{j=1}^n I(X_j \in B_i,\ |\hat{g}_n(X_j)| > r)\\
&\quad\geqslant \frac{1}{n}\sum_{j=1}^n I(X_j \in B_i)\\
&\qquad - \frac{1}{n}\sum_{j=1}^n I(X_j \in B_i,\ |\hat{g}_n(X_j)| > r)w(X_j)\Big/ \inf_{x\in B_i} w(x)\\
&\quad\geqslant \frac{1}{n}\sum_{j=1}^n I(X_j \in B_i) - \frac{1}{\inf_{x\in B_i} w(x)}\\
&\qquad \cdot \frac{1}{n}\sum_{j=1}^n I(X_j \in B_i,\ |\hat{g}_n(X_j)| > r)w(X_j). \qquad (14)
\end{aligned}$$

However, Markov's inequality and (12) imply that

$$\begin{aligned}
&\frac{1}{n}\sum_{j=1}^n I(|\hat{g}_n(X_j)| > r)w(X_j)\\
&\quad\leqslant r^{-2}\cdot\frac{1}{n}\sum_{j=1}^n \hat{g}_n(X_j)^2 w(X_j) \leqslant r^{-2}\beta \qquad (15)
\end{aligned}$$

for n sufficiently large. Choose r_0 so large that $r_0^{-2}\beta \leqslant \gamma/2 \min_{0\leqslant i\leqslant d+1} \inf_{x\in B_i} w(x)$. It follows from (14) and (15) that

$$\frac{1}{n}\sum_{j=1}^n I(X_j \in B_i,\ |\hat{g}_n(X_j)| \leqslant r_0) \geqslant \gamma/2 \qquad (16)$$

for n sufficiently large. For each such n, there therefore exists $X_{I(i)} \in B_i$ with $1 \leqslant I(i) \leqslant n$ and $|\hat{g}_n(X_{I(i)})| \leqslant r_0$. 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

$$\hat{g}_n(x) \leqslant \sum_{i=0}^d p_i \hat{g}_n(X_{I(i)}) \leqslant r_0. \qquad (17)$$

We now show that the convex hull contains $\tilde{\mathcal{H}} = \{x\colon \|x - (c/2d)e\| \leqslant c/4d\}$. We need to prove that there exist nonnegative $p_0, p_1, \ldots, p_d$ summing to 1 such that

$$\sum_{i=0}^d p_i X_{I(i)} = x$$

for $x \in \tilde{\mathcal{H}}$. This is equivalent to requiring a nonnegative solution $p_1, \ldots, p_d$ (summing to less than or equal to one) to the linear system

$$\sum_{i=1}^d p_i(ce_i + \tilde{X}_{I(i)} - X_{I(0)}) = x - X_{I(0)}, \qquad (18)$$

where $\tilde{X}_{I(i)} = X_{I(i)} - ce_i$ for $1 \leqslant i \leqslant d$. The linear system (18) can be reexpressed as

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

where $F = (F_{ij}\colon 1 \leqslant i, j \leqslant d)$ is a square $d \times d$ matrix in which the ith column is $\tilde{X}_{I(i)} - X_{I(0)}$. Set

$$|||F||| \triangleq \max_{1\leqslant i\leqslant d} \sum_{j=1}^d |F_{ij}|$$

and note that $|||c^{-1}F||| \leqslant \tau$. Hence, $(I + c^{-1}F)^{-1}$ exists and $|||(I + c^{-1}F)^{-1}||| \leqslant (1 - |||c^{-1}F|||)^{-1} \leqslant (1 - \tau)^{-1} \leqslant 2$. Therefore,

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

and hence

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

if $\|x\| \leqslant c$. It follows that if $x_i/c \geqslant 3\tau$ ($1 \leqslant i \leqslant d$) and $c^{-1}\sum_{i=1}^d x_i \leqslant 1 - 3\tau d$, the p_is solving (18) are nonnegative and sum to less than or equal to 1, so that $x = (x_1, \ldots, x_d)^T$ is in the convex hull of the $X_{I(i)}$s. If we choose τ so that $\tau \leqslant (12d)^{-1}$, it is easily seen that each $x \in \tilde{\mathcal{H}}$ satisfies these

linear equalities, so that $\tilde{\mathcal{H}}$ is contained in the convex hull. Relation (15) then implies that

$$\sup_{x\in\tilde{\mathcal{H}}} \hat{g}_n(x) \leqslant r_0 \tag{19}$$

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

$$X_{I(d+1)} = 1/2x + 1/2y \tag{20}$$

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

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

so that for n large

$$\begin{aligned}\hat{g}_n(x) &\geqslant 2\hat{g}_n(X_{I(d+1)}) - \hat{g}_n(y)\\ &\geqslant 2\hat{g}_n(X_{I(d+1)}) - r_0\\ &\geqslant -3r_0.\end{aligned}$$

Therefore, the lower bound ensues if we can write x as in (20). Observe that

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

so that

$$\begin{aligned}\|y - (c/2d)e\| &\leqslant \tau c/4d + \|x - (c/2d)e\|\\ &\leqslant c/8d + \|x - (c/2d)e\|.\end{aligned}$$

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

$$\inf_{x\in\mathcal{H}} \hat{g}_n(x) \geqslant -3r_0. \tag{21}$$

Combining (19) and (21) proves the result. □

Step 5. Set $\mathcal{H}_c = \{x\colon \|x\| \leqslant c\}$. 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,

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

for $x, y \in \mathcal{H}_c$ for n sufficiently large; see, for example, Van der Vaart and Wellner (1996, p. 165) and Roberts and Varberg (1974).

Step 6. Let

$$\begin{aligned}\mathcal{C}_c = \{&h\colon \mathcal{H}_c \to \mathbb{R} \text{ such that } h \text{ is convex on } \mathcal{H}_c,\\ &|h(x)| \leqslant \tilde{\beta}(c),\ |h(x) - h(y)| \leqslant 2/(c\delta)\tilde{\beta}(c(1+\delta))\|x - y\|\\ &\qquad\qquad \text{for } x, y \in \mathcal{H}_c\}\end{aligned}$$

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

$$d_c(h_1, h_2) = \sup_{x\in\mathcal{H}_c} |h_1(x) - h_2(x)|.$$

It follows that for each $\epsilon > 0$, there exists a finite collection of convex functions $h_1, h_2, \ldots, h_m$ such that $\bigcup_{i=1}^m \{h \in \mathcal{C}_c\colon d_c(h_i, h) < \epsilon\} \supseteq \mathcal{C}_c$ (i.e., $h_1, h_2, \ldots, h_m$ is an ϵ-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. We now use the empirical inner product inequality (10), and the fact that $\hat{g}_n$ can be uniformly approximated to precision ϵ within $\mathcal{H}_c$, to conclude that

$$\frac{1}{n}\sum_{i=1}^n (\hat{g}_n(X_i) - g_*(X_i))^2 w(X_i) \to 0 \tag{22}$$

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

To fill in the details, fix $\epsilon > 0$, choose c so large that the right-hand side of (13) is less than ϵ, and select $h_1, h_2, \ldots, h_m$ as suggested in Step 6. Then, for each $j \in \{1, \ldots, m\}$,

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^n (Y_i - g_*(X_i))(\hat{g}_n(X_i) - g_*(X_i))w(X_i)I(\|X_i\| \leqslant c)\\ &\quad\leqslant \frac{1}{n}\sum_{i=1}^n (Y_i - g_*(X_i))(\hat{g}_n(X_i) - h_j(X_i))w(X_i)I(\|X_i\| \leqslant c)\\ &\qquad + \frac{1}{n}\sum_{i=1}^n (Y_i - g_*(X_i))(h_j(X_i) - g_*(X_i))w(X_i)I(\|X_i\| \leqslant c)\\ &\quad\leqslant \frac{1}{n}\sum_{i=1}^n |Y_i - g_*(X_i)| \cdot \sup_{x\in\mathcal{H}_c} |\hat{g}_n(x) - h_j(x)| w(X_i)I(\|X_i\| \leqslant c)\\ &\qquad + \max_{1\leqslant r\leqslant m} \frac{1}{n}\sum_{i=1}^n (Y_i - g_*(X_i))(h_r(X_i)\\ &\qquad\qquad - g_*(X_i))w(X_i)I(\|X_i\| \leqslant c).\end{aligned}$$

As a consequence,

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^n (Y_i - g_*(X_i))(\hat{g}_n(X_i) - g_*(X_i))w(X_i)I(\|X_i\| \leqslant c)\\ &\quad\leqslant \min_{1\leqslant j\leqslant m} \sup_{x\in\mathcal{H}_c} |\hat{g}_n(x) - h_j(x)|\\ &\qquad \cdot \frac{1}{n}\sum_{i=1}^n |Y_i - g_*(X_i)| w(X_i)I(\|X_i\| \leqslant c)\\ &\qquad + \max_{1\leqslant r\leqslant m} \frac{1}{n}\sum_{i=1}^n (Y_i - g_*(X_i))(h_r(X_i)\\ &\qquad\qquad - g_*(X_i))w(X_i)I(\|X_i\| \leqslant c)\\ &\quad\leqslant \epsilon \cdot \frac{1}{n}\sum_{i=1}^n |Y_i - g_*(X_i)| w(X_i)\\ &\qquad + \max_{1\leqslant r\leqslant m} \frac{1}{n}\sum_{i=1}^n (Y_i - g_*(X_i))(h_r(X_i)\\ &\qquad\qquad - g_*(X_i))w(X_i)I(\|X_i\| \leqslant c),\end{aligned}$$

where we used the fact that $h_1, \ldots, h_m$ is an ϵ-net for $\mathcal{C}_c$ in the last inequality. Because $Y_i - g_*(X_i) = \nu_i$ and the h_i's are bounded, the strong law of large numbers for i.i.d. sequences implies that for $1 \leqslant r \leqslant m$,

$$\frac{1}{n}\sum_{i=1}^n \nu_i (h_r(X_i) - g_*(X_i))w(X_i)I(\|X_i\| \leqslant c) \to 0$$

a.s. as $n \to \infty$. It follows that

$$\limsup_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n}(Y_i - g_*(X_i))(\hat{g}_n(X_i) - g_*(X_i))w(X_i)I(\|X_i\| \leqslant c) \leqslant \epsilon\mathbb{E}[|Y_1 - g_*(X_1)|w(X_1)].$$

In view of Step 3, we therefore conclude that

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

yielding (22).

Step 8. We now use Step 7 and the fact that $\hat{g}_n$ is Lipschitz (uniformly in n) over each compact set $\mathcal{H}_c$ to conclude that $\hat{g}_n$ converges to g_* uniformly a.s. over $\mathcal{H}_c$.

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

$$\begin{aligned}|\hat{g}_n(x) - g_*(x)| &\leqslant |\hat{g}_n(x) - \hat{g}_n(X_i)| + |\hat{g}_n(X_i) - g_*(X_i)| \\ &\quad + |g_*(X_i) - g_*(x)| \\ &\leqslant \lambda(c)\epsilon + |\hat{g}_n(X_i) - g_*(X_i)| + \lambda(c)\epsilon,\end{aligned}$$

so

$$\begin{aligned}&\sup_{x\in B_j}|\hat{g}_n(x) - g_n(x)| \\ &\quad\leqslant 2\lambda(c)\epsilon + \frac{\sum_{i=1}^{n}|\hat{g}_n(X_i) - g_*(X_i)|I(X_i \in \Lambda_j)}{\sum_{i=1}^{n} I(X_i \in \Lambda_j)} \\ &\quad\leqslant 2\lambda(c)\epsilon + \frac{1}{n}\sum_{i=1}^{n}|\hat{g}_n(X_i) - g_*(X_i)|w(X_i)\cdot\frac{n}{\sum_{i=1}^{n} I(X_i \in \Lambda_j)} \\ &\quad\leqslant 2\lambda(c)\epsilon + \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{g}_n(X_i) - g_*(X_i))^2 w(X_i)}\sqrt{\frac{1}{n}\sum_{i=1}^{n} w(X_i)} \\ &\qquad\cdot\frac{n}{\sum_{i=1}^{n} I(X_i \in \Lambda_j)}.\end{aligned}$$

Taking advantage of Step 7, we conclude that

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

Because ϵ was arbitrary and there are only finitely many Λ_js, we conclude that

$$\sup_{x\in\mathcal{H}_c}|\hat{g}_n(x) - g_*(x)| \to 0 \quad \text{a.s.}$$

as $n \to \infty$, proving Theorem 1. □

Proof of Theorem 2. The proof of Theorem 2 is identical to that of Theorem 1 except that Steps 6 and 7 are replaced by Steps 6′ and 7′. (Note that Steps 4 and 5 follow immediately from the fact that $0 \leqslant \hat{g}_n(x) \leqslant k(x)$ for $x \in \mathbb{R}^d$.)

Step 6′. Let

$$\begin{aligned}\mathcal{C}_c' = \Big\{&h\colon \mathcal{H}_c \to \mathbb{R} \text{ such that there is a convex function } \tilde{h}\colon \\ &\mathbb{R}^d \to \mathbb{R} \text{ agreeing with } h \text{ on } \mathcal{H}_c, \\ &|\tilde{h}(x)| \leqslant \tilde{\beta}(c(1+\delta)) \text{ for } x \in \mathcal{H}_{c(1+\delta)}, \\ &|\tilde{h}(x) - \tilde{h}(y)| \leqslant 2/(c\delta)\tilde{\beta}(c(1+\delta))\|x - y\| \\ &\qquad\qquad \text{for } x, y \in \mathcal{H}_c, \\ &0 \leqslant \tilde{h}(x) \leqslant k(x) \text{ for } x \in \mathbb{R}^d\Big\}\end{aligned}$$

and note that Steps 4 and 5 guarantee that for each $c \geqslant 0$, $\hat{g}_n$ restricted on $\mathcal{H}_c$ belongs to $\mathcal{C}_c'$ for n sufficiently large a.s. Furthermore, $\mathcal{C}_c'$ is a closed subset of $\mathcal{C}_c$ in the uniform metric d_c given by

$$d_c(h_1, h_2) = \sup_{x\in\mathcal{H}_c}|h_1(x) - h_2(x)|.$$

To see why $\mathcal{C}_c'$ is closed, let $(h_i\colon i \geqslant 1)$ be a convergent sequence in $\mathcal{C}_c$ with limit $h_\infty \in \mathcal{C}_c$ for which $h_i \in \mathcal{C}_c'$ for $1 \leqslant i < \infty$. For $1 \leqslant i < \infty$, there exists an extension $\tilde{h}_i$ of h_i to $\mathbb{R}^d$ that is nonnegative, convex, and bounded above by k. Set $\tilde{h}_\infty = \limsup \tilde{h}_k$ and note that $\tilde{h}_\infty = h_\infty$ on $\mathcal{H}_c$. Furthermore, $\tilde{h}_\infty$ is convex, nonnegative, and bounded above by k because these properties are inherited under pointwise convergence.

Because $\mathcal{C}_c'$ is a closed subset of $\mathcal{C}_c$ and $\mathcal{C}_c$ is compact, $\mathcal{C}_c'$ is compact by Theorem 26.2 of Munkres (2000). It follows that for each $\epsilon > 0$, there exists a finite collection of convex functions $h_1, h_2, \ldots, h_m$ such that $\bigcup_{i=1}^{m}\{h \in \mathcal{C}_c'\colon d_c(h_i, h) < \epsilon\} \supseteq \mathcal{C}_c'$, and there exists a convex function $\tilde{h}_i\colon \mathbb{R}^d \to \mathbb{R}$ that agrees with h_i on $\mathcal{H}_c$ and $0 \leqslant \tilde{h}_i(x) \leqslant k(x)$ for $x \in \mathbb{R}^d$ for $1 \leqslant i \leqslant m$.

Step 7′. We now use the empirical inner product inequality (10), and the fact that $\hat{g}_n$ can be uniformly approximated to precision ϵ within $\mathcal{H}_c$, to conclude that

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{g}_n(X_i) - g_*(X_i))^2 w(X_i) \to 0 \tag{23}$$

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

To fill in the details, fix $\epsilon > 0$, choose c so large that the right-hand side of (13) is less than ϵ and $\mathbb{E}[(Y - g_*(X))^2 w(X) I(\|X\| > c)] < \epsilon$, and select $\tilde{h}_1, \tilde{h}_2, \ldots, \tilde{h}_m$ as suggested in Step 6′. Then, for each $j \in \{1, \ldots, m\}$,

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}(Y_i - g_*(X_i))(\hat{g}_n(X_i) - g_*(X_i))w(X_i)I(\|X_i\| \leqslant c) \\ &\leqslant \frac{1}{n}\sum_{i=1}^{n}(Y_i - g_*(X_i))(\hat{g}_n(X_i) - \tilde{h}_j(X_i))w(X_i)I(\|X_i\| \leqslant c) \\ &\quad + \frac{1}{n}\sum_{i=1}^{n}(Y_i - g_*(X_i))(\tilde{h}_j(X_i) - g_*(X_i))w(X_i) \\ &\quad + \frac{1}{n}\sum_{i=1}^{n}|Y_i - g_*(X_i)||\tilde{h}_j(X_i) - g_*(X_i)|w(X_i)I(\|X_i\| > c)\end{aligned}$$

$$\leqslant \frac{1}{n}\sum_{i=1}^{n}|Y_i - g_*(X_i)| \cdot \sup_{x\in\mathcal{H}_c}|\hat{g}_n(x) - \tilde{h}_j(x)|w(X_i)I(\|X_i\| \leqslant c)$$
$$+ \max_{1\leqslant r\leqslant m}\frac{1}{n}\sum_{i=1}^{n}(Y_i - g_*(X_i))(\tilde{h}_r(X_i) - g_*(X_i))w(X_i)$$
$$+ \max_{1\leqslant r\leqslant m}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - g_*(X_i))^2 w(X_i)I(\|X_i\| > c)}$$
$$\cdot\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\tilde{h}_r(X_i) - g_*(X_i))^2 w(X_i)}.$$

As a consequence,

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - g_*(X_i))(\hat{g}_n(X_i) - g_*(X_i))w(X_i)I(\|X_i\| \leqslant c)$$
$$\leqslant \min_{1\leqslant j\leqslant m}\sup_{x\in\mathcal{H}_c}|\hat{g}_n(x) - \tilde{h}_j(x)|$$
$$\cdot\frac{1}{n}\sum_{i=1}^{n}|Y_i - g_*(X_i)|w(X_i)I(\|X_i\| \leqslant c)$$
$$+ \max_{1\leqslant r\leqslant m}\frac{1}{n}\sum_{i=1}^{n}(Y_i - g_*(X_i))(\tilde{h}_r(X_i) - g_*(X_i))w(X_i)$$
$$+ \max_{1\leqslant r\leqslant m}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - g_*(X_i))^2 w(X_i)I(\|X_i\| > c)}$$
$$\cdot\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\tilde{h}_r(X_i) - g_*(X_i))^2 w(X_i)}$$
$$\leqslant \epsilon\cdot\frac{1}{n}\sum_{i=1}^{n}|Y_i - g_*(X_i)|w(X_i)$$
$$+ \max_{1\leqslant r\leqslant m}\frac{1}{n}\sum_{i=1}^{n}(Y_i - g_*(X_i))(\tilde{h}_r(X_i) - g_*(X_i))w(X_i)$$
$$+ \max_{1\leqslant r\leqslant m}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - g_*(X_i))^2 w(X_i)I(\|X_i\| > c)}$$
$$\cdot\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\tilde{h}_r(X_i) - g_*(X_i))^2 w(X_i)},$$

where we used the fact that $h_1, \ldots, h_m$ is an ϵ-net for $\mathcal{C}'_c$ in the last inequality. Because $\tilde{h}_1, \ldots, \tilde{h}_m$ are in $L^2(X)$, the strong law of large numbers for i.i.d. sequences guarantees that

$$\limsup_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}(Y_i - g_*(X_i))(\hat{g}_n(X_i) - g_*(X_i))w(X_i)I(\|X_i\| \leqslant c)$$
$$\leqslant \epsilon\mathbb{E}[|Y_1 - g_*(X_1)|w(X_1)]$$
$$+ \max_{1\leqslant r\leqslant m}\mathbb{E}[(Y_1 - g_*(X_1))(\tilde{h}_r(X_1) - g_*(X_1))w(X_1)]$$
$$+ \max_{1\leqslant r\leqslant m}\sqrt{\mathbb{E}[(Y_1 - g_*(X_1))^2 w(X_1)I(\|X_1\| > c)]}$$
$$\cdot\sqrt{\mathbb{E}[(\tilde{h}_r(X_1) - g_*(X_1))^2 w(X_1)]}$$
$$\leqslant \epsilon\sqrt{\mathbb{E}[(Y_1 - g_*(X_1))^2 w(X_1)]}\cdot\sqrt{\mathbb{E}[w(X_1)]}$$
$$+ \max_{1\leqslant r\leqslant m}\langle f_* - g_*, \tilde{h}_r - g_*\rangle$$
$$+ \epsilon\max_{1\leqslant r\leqslant m}\sqrt{\mathbb{E}[(\tilde{h}_r(X_1) - g_*(X_1))^2 w(X_1)]}$$

a.s. Because $\tilde{h}_r \in \tilde{\mathcal{C}}^2$ for $1 \leqslant r \leqslant m$, it follows that $\langle f_* - g_*, \tilde{h}_r - g_*\rangle \leqslant 0$ for $1 \leqslant r \leqslant m$. In view of Step 3, we therefore conclude that

$$\limsup_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}(Y_i - g_*(X_i))(\hat{g}_n(X_i) - g_*(X_i))w(X_i)$$
$$\leqslant \epsilon\sqrt{\mathbb{E}[(Y_1 - g_*(X_1))^2 w(X_1)]}\cdot\sqrt{\mathbb{E}[w(X_1)]}$$
$$+ \epsilon\max_{1\leqslant r\leqslant m}\sqrt{\mathbb{E}\left[(\tilde{h}_r(X_1) - g_*(X_1))^2 w(X_1)\right]} + \epsilon$$

a.s. Because $(a+b)^2 \leqslant 2a^2 + 2b^2$ for $a, b \in \mathbb{R}$, and $0 \leqslant \tilde{h}_r(x), g_*(x) \leqslant k(x)$ for $x \in \mathbb{R}^d$, it follows

$$\limsup_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}(Y_i - g_*(X_i))(\hat{g}_n(X_i) - g_*(X_i))w(X_i)$$
$$\leqslant \epsilon\sqrt{\mathbb{E}[(Y_1 - g_*(X_1))^2 w(X_1)]}\cdot\sqrt{\mathbb{E}[w(X_1)]}$$
$$+ \epsilon\sqrt{4\mathbb{E}[k(X_1)^2 w(X_1)]} + \epsilon.$$

Because $\epsilon > 0$ was arbitrary, (10) implies that

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

yielding (23).

4. Extensions

We consider here two extensions of the convex regression problem discussed in §§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 continues 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 $(1/2)e$, 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 §3.

Extension 2: Adding the requirement that the function to be estimated be both convex and nondecreasing.

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 nondecreasing. In particular, we say that a function $g\colon \mathbb{R}^d \to \mathbb{R}$ is *nondecreasing* if $g(x) \leqslant g(y)$ whenever $x \leqslant y$ (so that $x_i \leqslant y_i$ for $1 \leqslant i \leqslant d$). We now adapt the definition of the cone of functions $\mathscr{C}$ to $\mathscr{C} = \{g\colon \mathbb{R}^d \to \mathbb{R}$ such that g is convex and nondecreasing$\}$, and assume A1–A4. We can now analogously define L^2 and $\mathscr{C}^2$ as in §2.

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

$$\begin{aligned}\min\ & \frac{1}{n}\sum_{i=1}^{n} w(X_i)(Y_i - g_i)^2\\ \text{s/t}\ & g_j \geqslant g_i + \xi_i^T(X_j - X_i), \quad 1 \leqslant i, j \leqslant n\\ & \xi_i \geqslant 0, \quad 1 \leqslant i \leqslant n.\end{aligned}$$

As in §2, this defines $\hat{g}_n(\cdot)$ at the X_is. To define $\hat{g}_n(\cdot)$ globally, we again define $\hat{g}_n(\cdot)$ via (4) (but with our modified definition of $\mathscr{C}$); it is easily seen that $\hat{g}_n$ is convex and nondecreasing on $\mathrm{conv}(X_1, \ldots, X_n)$ (because the supremum of convex nondecreasing functions is necessarily convex and nondecreasing). In addition, for $x \neq X_i$, $1 \leqslant i \leqslant n$, $\hat{g}_n(x)$ can be evaluated by solving an LP:

$$\begin{aligned}\max\ & y\\ \text{s/t}\ & \hat{g}_j \geqslant \hat{g}_i + \xi_i^T(X_j - X_i), \quad 1 \leqslant i, j \leqslant n\\ & y \geqslant \hat{g}_i + \xi_i^T(x - X_i), \quad 1 \leqslant i \leqslant n\\ & \hat{g}_j \geqslant y + \tilde{\xi}^T(X_j - x), \quad 1 \leqslant j \leqslant n\\ & \tilde{\xi} \geqslant 0, \quad \xi_i \geqslant 0, \quad 1 \leqslant i \leqslant n.\end{aligned}$$

Because $\mathscr{C}$ is a subcone of the cone of convex functions used in §3, Steps 1 through 5 of §3 follow as in that section. For Step 6, modify $\mathscr{C}_c$ so that the requirement that h is nondecreasing is added to the definition provided in §3. Clearly, $\mathscr{C}_c$ is a bounded and equicontinuous family of functions. Furthermore, any pointwise limit of functions lying in $\mathscr{C}_c$ must be convex, nondecreasing, and satisfy the Lipschitz and boundedness constraints, so that $\mathscr{C}_c$ is closed. It follows that $\mathscr{C}_c$ 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 ϵ-net $h_1, h_2, \ldots, h_m$ of elements in $\mathscr{C}_c$ for which their corresponding ϵ-balls cover $\mathscr{C}_c$.

Steps 7 and 8 follow the same lines as in §3, so that we therefore obtain the following theorem.

THEOREM 3. *Assume* A1–A4 *and that* $f_* \in \mathscr{C}$. *Then, for each* $c \geqslant 0$,

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

as $n \to \infty$.

5. Numerical Results

In this section, we numerically investigate the performance of our least-squares estimator, and compare it to the performance of conventional linear regression and the competing maximum-likelihood estimator of Allon et al. (2007) (which they term the *convex entropy nonparametric* (CENP) estimator). Note that unless one explicitly includes nonlinear functions as explanatory (independent) variables in one's linear regression (e.g., regress on terms of the form $x_i^j x_k^l$, in addition to $x_1, x_2, \ldots, x_d$), it is clear that linear regression will not be able to accurately approximate highly nonlinear functions. Thus, the success of linear regression (in a setting in which the underlying function to be approximated is convex) depends largely on the explanatory variables that one chooses to include. Given the enormous attitude that the user has in this regard, this can make use of linear regression in the nonlinear setting very challenging. In our experiments, we will only use linear terms $x_1, \ldots, x_d$ to compute the regression equation as follows. Let $\hat{a}_n(0), \ldots, \hat{a}_n(d) \in \mathbb{R}$ be the solution to

$$\min_{a_0, a_1, \ldots, a_d \in \mathbb{R}} \sum_{i=1}^{n} (Y_i - (a_0 + a_1 X_1^i + \cdots + a_d X_d^i))^2,$$

where $X_i = (X_1^i, \ldots, X_d^i)$. For $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$, the linear regression estimator $\hat{l}_n(x)$ is computed by $\hat{l}_n(x) = \hat{a}_n(0) + \hat{a}_n(1)x_1 + \cdots + \hat{a}_n(d)x_d$.

As noted in the introduction, the CENP estimator concerns a multiplicative noise model, namely,

$$Y_i = f_*(X_i)\eta_i, \quad 1 \leqslant i \leqslant n, \tag{24}$$

where $f_*\colon \mathbb{R}^d \to \mathbb{R}$ is assumed to be nondecreasing and concave with $f_*(0) = 0$, and the η_is are i.i.d. positive random variables independent of the X_is. In particular, the CENP estimator is defined as the maximizer, over nondecreasing concave f with $f(0) = 0$, of the (normalized) log-likelihood

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n} \log k\left(\log\left(\frac{Y_i}{f(X_i)}\right)\right)\\ &\quad \triangleq -\frac{1}{n}\sum_{i=1}^{n} h\left(\log\eta_i + \log\left(\frac{f_*(X_i)}{f(X_i)}\right)\right),\end{aligned} \tag{25}$$

where $h(x) = -\log k(x)$ and k is the density of $\log\eta_1$. It is further assumed in Allon et al. (2007) that the density k is even and log-concave (so that $\log k$ is concave). By taking logarithms in (24), the model is transformed to an "additive

error" model that clearly resembles the model (1) associated with our least-squares estimator. In particular,

$$\tilde{Y}_i = \tilde{f}_*(X_i) + \nu_i, \tag{26}$$

where $\tilde{Y}_i = \log Y_i$, $\tilde{f}_* = \log f_*$, and $\nu_i = \log \eta_i$; the function $\tilde{f}_*$ is nondecreasing and concave (because concavity is preserved under a logarithm transformation). Conversely, if we start with the model (26) and assume that $\exp(\tilde{f}_*(\cdot))$ is concave (which is a stronger hypothesis than assuming that $\tilde{f}_i$ is itself concave) and nondecreasing, we arrive at the model (24). Thus, the CENP estimator, which was introduced in the setting of the multiplicative model (24), has an additive error interpretation.

In this additive error context, the assumptions on both the function (concavity of the exponential of the function versus just concavity of the function itself) and the noise (a symmetric log-concave density with the same distribution asross all x-values versus just an assumption of zero expectation) are much stronger than those made for (1). Of course, it is in principle possible that these stronger assumptions are unnecessary to guarantee consistency and that the CENP estimator is consistent for the additive error model (26) under the same conditions as is our least-squares estimator. We now show that this is false, and that the least-squares estimator does indeed converge under (much) weaker conditions than does the CENP estimator. Thus, the least-squares estimator analyzed in this paper is (much) more robust than is the CENP estimator for the additive model.

Suppose that $k(x) = 1/2\exp(-|x|)$ so that $h(x) = |x| + \log 2$. However, assume that the modeler has misspecified the density of the noise, so that the $\log \eta_i$s actually come from an i.i.d. sample associated with another positive density $\tilde{k}$. The law of large numbers ensures that the normalized log-likelihood (25) converges a.s. to $-\log 2 - \mathbb{E}[|\log \eta_1 - a|]$ at $f(\cdot) = e^{-a} f_*(\cdot)$. It is well known that the a that maximizes this expression is $a^* =$ median of $\log \eta_1$. Of course, when $\tilde{k} = k$, $a^* = 0$ (because k is even) so that the maximizing f (over f of the form $f = e^{-a} f_*$) is $f = f_*$, as desired. However, if $\tilde{k}$ is actually nonsymmetric, the median a^* of $\log \eta_1$ is nonzero, so that the maximizing f is no longer f_*. In this case, the CENP estimator will be inconsistent (whereas the least-squares estimator will be consistent so long as the mean of $\tilde{k}$ equals zero).

Given that we have posed the question of whether the CENP estimator, originally derived for the multiplicative model (24), is consistent for the additive model under conditions as general as those associated with our least-squares estimator, we can ask a related question: How does our least-squares estimator behave if the actual error model is multiplicative rather than additive? Suppose that the data are generated by the model (24), and that the (X_i, Y_i)'s are i.i.d. with $\mathbb{E}[Y_i \mid X_i] = f_*(X_i)$. For each convex f, the least-squares estimator converges a.s. to

$$\begin{aligned}
&\mathbb{E}[(Y_1 - f(X_1))^2 w(X_1)] \\
&\quad = \mathbb{E}[(\eta_1 f(X_1) - f(X_1))^2 w(X_1)] \\
&\quad = \mathbb{E}[((\eta_1 - 1) f_*(X_1) + (f_*(X_1) - f(X_1)))^2 w(X_1)] \\
&\quad = \mathbb{E}[(\eta_1 - 1)^2 f_*(X_1)^2 w(X_1)] \\
&\qquad + 2\mathbb{E}[(\eta_1 - 1) f_*(X_1)(f_*(X_1) - f(X_1)) w(X_1)] \\
&\qquad + \mathbb{E}[(f_*(X_1) - f(X_1))^2 w(X_1)].
\end{aligned}$$

However, $\mathbb{E}[(\eta_1 - 1) f_*(X_1)(f_*(X_1) - f(X_1)) w(X_1) | X_1] = 0$, so that $\mathbb{E}[(Y_1 - f(X_1))^2 w(X_1)]$ is minimized by $f = f_*$ when f_* is convex. Thus, although the present paper does not provide complete rigorous details, we can expect that a proof similar to that provided in the current paper establishes a.s. consistency of our least-squares estimator for the multiplicative model, again under weaker conditions than for the CENP estimator (because no assumption of symmetry or log-concavity for the density of $\log \eta_1$ is necessary).

Having discussed some of the differences between our estimator, the CENP estimator, and (conventional) linear regression, we devote the remainder of this section to our computational examples.

Example 1 (Consumer Preference Function). Suppose we wish to predict consumer preferences by estimating a consumer utility function. The underlying utility function $u\colon \mathbb{R} \to \mathbb{R}$ is given by

$$u(x) = 5 - 5\exp(-2x) \tag{27}$$

for $x \geqslant 0$. Now suppose we are given some consumer preference data $(X_1, Y_1), \ldots, (X_n, Y_n)$, where $X_i = i/n$ for $1 \leqslant i \leqslant n$ and the noisy measurement Y_i of $u(X_i)$ is given by $u(X_i) + (\log N_i(0,1) - e^{1/2})/5$, and $\log N_i(0,1)$ is i.i.d. and follows a lognormal distribution with normal mean 0 and variance 1. Table 1 reports the averages (Mean) and the standard deviation (Std) of the linear regression estimator, the CENP estimator, and our proposed estimator $\hat{g}_n$ at $X_i = 0.5$ based on 100 replications for each value of n. In this example, as in the others of this section, the CENP estimator that we implement uses the choice $h(x) = \exp(x) + \exp(-x)$.

Example 2 (Single-Server Queue). We consider a single-server queue with infinite buffer, where the interarrival times follow a lognormal distribution with normal

Table 1. Performance of the linear regression, the CENP, and $\hat{g}_n(\cdot)$ as estimators for $u(\cdot)$ in (27).

	$X_i = 0.5$					
	Linear regression		CENP		$\hat{g}_n(X_i)$	
n	Mean	Std	Mean	Std	Mean	Std
16	2.86	0.11	3.13	0.18	3.17	0.21
32	2.84	0.08	3.13	0.13	3.16	0.14
64	2.84	0.05	3.13	0.12	3.16	0.11
128	2.84	0.04	3.13	0.11	3.16	0.08
$u(X_i)$			3.16			

mean $(\ln 1.5) - 0.5$ and variance 1, and the service times follow a lognormal distribution with normal mean $(\ln x) - 0.5$ and variance 1. The interarrival times and service times are independent, and the service discipline is first in/first out (FIFO). We wish to compute the expected waiting time w_{100} of the 100th customer as a function of the mean service time x. Even though there is no explicit formula for $w_{100}(x)$, the convexity of $w_{100}(\cdot)$ has been theoretically established; see Shanthikumar and Yao (1991, p. 137) for details. To compute w_{100}, we simulated the single-server queue at $x = X_i = 0.5 + i/(2n)$ for $1 \leqslant i \leqslant n$, and computed the waiting time Y_i of the 100th customer. We then obtained the average of 10 independent copies of Y_i, say $\bar{Y}_i$. Using $(X_1, \bar{Y}_1), \ldots, (X_n, \bar{Y}_n)$, we computed the linear regression estimator, the CENP estimator, and our proposed estimator $\hat{g}_n$ as the estimators for w_{100}.

Table 2 reports the averages (Mean) and the standard deviation (Std) of the linear regression estimator, the CENP estimator, and our proposed estimator $\hat{g}_n$ at $X_i = 0.75$ based on 100 replications for each value of n. The value of $w_{100}(X_i)$ in the last row of Table 2 is obtained from averaging 100,000 independent copies of Y_i at X_i.

Example 3 (Tandem Queue). We consider a system of two single-server stations in tandem, where the interarrival times follow a lognormal distribution with normal mean $(\ln 3) - 0.5$ and variance 1, and the service times at server k follow a lognormal distribution with normal mean $(\ln x^k) - 0.5$ and variance 1 for $k = 1, 2$. The interarrival times and service times are independent and the service discipline is FIFO at each server. Each server has unlimited buffer space. We wish to compute the expected sojourn time $s_5(x^1, x^2)$ of the fifth customer. Again, even though there is no explicit formula for $s_5(\cdot, \cdot)$, the convexity of s_5 has been proven; see p. 141 of Shanthikumar and Yao (1991) for details. To compute s_5, we simulated the tandem queue at $(x^1, x^2) = X_{ij} = (1.1 + i/(4n^{1/2}), 1.1 + j/(4n^{1/2}))$ for $1 \leqslant i, j \leqslant n^{1/2}$ and computed the sojourn time Y_{ij} of the fifth customer. We then obtained the average of 10 independent copies of Y_{ij}, say $\bar{Y}_{ij}$. Using $(X_{ij}, \bar{Y}_{i,j})$ for $1 \leqslant i, j \leqslant n$, we computed the linear regression estimator, the CENP estimator, and our proposed estimator $\hat{g}_n$ as the estimators for s_5.

Table 2. Performance of the linear regression, the CENP, and $\hat{g}_n(\cdot)$ as estimators for $w_{100}(\cdot)$.

	$X_i = 0.75$					
	Linear regression		CENP		$\hat{g}_n(X_i)$	
n	Mean	Std	Mean	Std	Mean	Std
16	1.32	0.21	0.91	0.26	1.09	0.25
32	1.32	0.15	0.90	0.22	1.11	0.18
64	1.32	0.11	0.87	0.15	1.12	0.13
128	1.32	0.08	0.86	0.15	1.13	0.11
256	1.33	0.06	0.86	0.15	1.13	0.07
$w_{100}(X_i)$			1.13			

Table 3. Performance of the linear regression, the CENP, and $\hat{g}_n(\cdot)$ as estimators for $s_5(\cdot)$.

	$X_{ij} = (1.225, 1.225)$					
	Linear regression		CENP		$\hat{g}_n(X_i)$	
n	Mean	Std	Mean	Std	Mean	Std
36	3.88	0.26	3.90	0.21	4.26	0.22
64	4.00	0.20	3.96	0.18	4.28	0.16
100	4.05	0.19	3.99	0.15	4.28	0.14
144	4.09	0.14	4.00	0.10	4.28	0.12
$s_5(X_{ij})$			4.28			

Table 3 reports the averages (Mean) and the standard deviation (Std) of the linear regression estimator, the CENP estimator, and our proposed estimator $\hat{g}_n$ at $X_{ij} = (1.225, 1.225)$ based on 100 replications for each value of n. The value of $s_5(X_{ij})$ in the last row of Table 3 is obtained from averaging 100,000 independent copies of Y_{ij} at X_{ij}.

Acknowledgments

The research of the second author was partially supported by a grant from the King Abdullah University of Science and Technology. The authors wish to express their thanks to the referees and associate editor for their suggestions and comments, which served to significantly improve the quality of the paper.

References

Allon, G., M. Beenstock, S. Hackman, U. Passy, A. Shapiro. 2007. Nonparametric estimation of concave production technologies by entropy. *J. Appl. Econometrics* **22** 795–816.

Avriel, M. 1976. *Nonlinear Programming: Analysis and Methods*. Prentice-Hall, Englewood Cliffs, NJ.

Barlow, R. E., H. D. Brunk. 1972. The isotonic regression problem and its dual. *J. Amer. Statist. Assoc.* **67**(337) 140–147.

Boyd, S., L. Vandenberghe. 2004. *Convex Optimization*. Cambridge University Press, Cambridge, UK.

Bronshtein, E. M. 1976. ϵ-entropy of convex sets and functions. *Siberian Math. J.* **17**(3) 393–398.

Brunk, H. D. 1958. On the estimation of parameters restricted by inequalities. *Ann. Math. Statist.* **29**(2) 437–454.

Brunk, H. D. 1965. Conditional expectations given a σ-lattice and applications. *Ann. Math. Statist.* **36**(5) 1339–1350.

Chen, H., D. D. Yao. 2001. *Fundamentals of Queueing Networks: Performance, Asymptotics, and Optimization*. Springer-Verlag, New York.

Copson, E. T. 1972. *Metric Spaces*. Cambridge University Press, London.

Cule, M. L., R. J. Samworth. 2010. Theoretical properties of the log-concave maximum likelihood estimator of a multidimensional density. *Electron. J. Statist.* **4** 254–270.

Dümbgen, L., R. J. Samworth, D. Schuhmacher. 2011. Approximation by log-concave distributions with applications to regression. *Ann. Statist.* **39**(2) 702–730.

Groeneboom, P., G. Jongbloed, J. A. Wellner. 2001. Estimation of a convex function: Characterizations and asymptotic theory. *Ann. Statist.* **29**(6) 1653–1698.

Hall, P., L. Huang. 2001. Nonparametric kernel regression subject to monotonicity constraints. *Ann. Statist.* **29**(3) 624–647.

Hanson, D. L., G. Pledger. 1976. Consistency in concave regression. *Ann. Statist.* **4**(6) 1038–1050.

Kuosmanen, T. 2008. Representation theorem for convex nonparametric least squares. *Econometrics J.* **11** 308–325.

Lim, E., P. W. Glynn. 2006. Simulation-based response surface computation in the presence of monotonicity. *Proc. 2006 Winter Simulation Conf., Monterey, CA*, 264–271.

Mammen, E. 1991. Nonparametric regression under qualitative smoothness assumptions. *Ann. Statist.* **19**(2) 741–759.

Meyer, R. F., J. W. Pratt. 1968. The consistent assessment and fairing of preference functions. *IEEE Trans. Systems Sci. Cybernetics* **4**(3) 270–278.

Munkres, J. R. 2000. *Topology*, 2nd ed. Prentice Hall, Upper Saddle River, NJ.

Roberts, A. W., D. E. Varberg. 1974. Another proof that convex functions are locally Lipschitz. *Amer. Math. Monthly* **81**(9) 1014–1016.

Rockafellar, R. T. 1974. *Conjugate Duality and Optimization*. Society for Industrial and Applied Mathematics, Philadelphia.

Rosenblatt, F. 1956. Remarks on some nonparametric estimates of a density function. *Ann. Statist.* **27**(3) 832–837.

Seijo, E., B. Sen. 2011. Nonparametric least squares estimation of a multivariate convex regression function. *Ann. Statist.* **39**(3) 1633–1657.

Shanthikumar, J. G., D. D. Yao. 1991. Strong stochastic convexity: Closure properties and applications. *J. Appl. Probab.* **28** 131–145.

Van der Vaart, A. W., J. A. Wellner. 1996. *Weak Convergence and Empirical Processes*. Springer–Verlag, New York.

Varian, H. R. 1984. The nonparametric approach to production analysis. *Econometrica* **52**(3) 579–597.

Varian, H. R. 1985. Nonparametric analysis of optimizing behavior with measurement error. *J. Econometrics* **30** 445–458.

Wright, F. T. 1979. A strong law for variables indexed by a partially ordered set with applications to isotone regression. *Ann. Probab.* **7**(1) 109–127.

Yatchew, A. J., L. Bos. 1997. Nonparametric least squares regression and testing in economic models. *J. Quant. Econom.* **13** 81–131.