The Annals of Statistics
2011, Vol. 39, No. 3, 1633–1657
DOI: 10.1214/10-AOS852

NONPARAMETRIC LEAST SQUARES ESTIMATION OF A MULTIVARIATE CONVEX REGRESSION FUNCTION

BY EMILIO SEIJO AND BODHISATTVA SEN[1]

Columbia University

This paper deals with the consistency of the nonparametric least squares estimator of a convex regression function when the predictor is multidimensional. We characterize and discuss the computation of such an estimator via the solution of certain quadratic and linear programs. Mild sufficient conditions for the consistency of this estimator and its subdifferentials in fixed and stochastic design regression settings are provided.

1. Introduction. Consider a closed, convex set $\mathcal{X} \subset \mathbb{R}^d$, for $d \geq 1$, with nonempty interior and a regression model of the form

$$Y = \phi(X) + \varepsilon, \tag{1}$$

where X is a $\mathcal{X}$-valued random vector, ε is a random variable with $\mathbf{E}(\varepsilon|X) = 0$, and $\phi : \mathbb{R}^d \to \mathbb{R}$ is an unknown *convex* function. Given independent observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ from such a model, we wish to estimate ϕ by the method of least squares, that is, by finding a convex function $\hat{\phi}_n$ which minimizes the discrete $\mathcal{L}_2$ norm

$$\left(\sum_{k=1}^{n} |Y_k - \psi(X_k)|^2 \right)^{1/2}$$

among all convex functions ψ defined on the convex hull of $X_1, \ldots, X_n$. In this paper we characterize the least squares estimator, provide means for its computation, study its finite sample properties and prove its consistency.

The problem just described is a nonparametric regression problem with known shape restriction (convexity). Such problems have a long history in the statistical literature with seminal papers like Brunk (1955), Grenander (1956) and Hildreth (1954) written more than 50 years ago, albeit in simpler settings. The former two papers deal with the estimation of monotone functions while the latter discusses least squares estimation of a concave function whose domain is a subset of the real line. Since then, many results on different nonparametric shape restricted regression problems have been published; see, for instance, Brunk (1970) and, more

Received February 2010; revised September 2010.

[1]Supported by NSF Grant DMS-09-06597.

MSC2010 subject classifications. Primary 62G08, 62G05; secondary 62G20.

Key words and phrases. Consistency, linear program, semidefinite quadratic program, shape restricted estimation, subdifferentials.

recently, Zhang (2002) for literature concerning isotonic regression. In the particular case of convex regression, Hanson and Pledger (1976) proved the consistency of the least squares estimator introduced in Hildreth (1954). Some years later, Mammen (1991) and Groeneboom, Jongbloed and Wellner (2001) derived, respectively, the rate of convergence and asymptotic distribution of this estimator. Some alternative methods of estimation that combine shape restrictions with smoothness assumptions have also been proposed for the one-dimensional case; see, for example, Birke and Dette (2006) where a kernel-based estimator is defined and its asymptotic distribution derived.

Although the asymptotic theory of the one-dimensional convex regression problem is well understood, not much has been done in the multidimensional scenario. The absence of a natural order structure in $\mathbb{R}^d$, for $d > 1$, poses a natural impediment in such extensions. A convex function on the real line can be characterized as an absolutely continuous function with increasing first derivative [see, e.g., Folland (1999), Exercise 42.b, page 109]. This characterization plays a key role in the computation and asymptotic theory of the least squares estimator in the one-dimensional case. By contrast, analogous results for convex functions of several variables involve more complicated characterizations using either second-order conditions [as in Dudley (1977), Theorem 3.1, page 163] or cyclical monotonicity [as in Rockafellar (1970), Theorems 24.8 and 24.9, pages 238 and 239]. Interesting differences between convex functions on $\mathbb{R}$ and convex functions on $\mathbb{R}^d$, for $d > 1$, are given in Johansen (1974) and Bronšteĭn (1978).

Recently there has been considerable interest in shape restricted function estimation in multidimension. In the density estimation context, Cule, Samworth and Stewart (2010) deal with the computation of the nonparametric maximum likelihood estimator of a multidimensional log-concave density, while Cule and Samworth (2010), Schuhmacher, Hüsler and Dümbgen (2009) and Schuhmacher and Dümbgen (2010) discuss its consistency and related issues. Seregin and Wellner (2009) study the computation and consistency of the maximum likelihood estimator of convex-transformed densities. This paper focuses on estimating a regression function which is known to be convex. To the best of our knowledge this is the first attempt to systematically study the characterization, computation and consistency of the least squares estimator of a convex regression function with multidimensional covariates in a *completely nonparametric* setting.

In the field of econometrics some work has been done on this multidimensional problem in less general contexts and with more stringent assumptions. Estimation of concave and/or componentwise nondecreasing functions has been treated, for example, in Banker and Maindiratta (1992), Matzkin (1991, 1993), Beresteanu (2007) and Allon et al. (2007). The first two papers define maximum likelihood estimators in semiparametric settings. The estimators in Matzkin (1991) and Banker and Maindiratta (1992) are shown to be consistent in Matzkin (1991) and Sarath and Maindiratta (1997), respectively. A maximum likelihood estimator and a sieved least squares estimator have been defined and techniques

for their computation have been provided in Allon et al. (2007) and Beresteanu (2007), respectively.

The method of least squares has been applied to multidimensional concave regression in Kuosmanen (2008). We take this work as our starting point. In agreement with the techniques used there, we define a least squares estimator which can be computed by solving a quadratic program. We argue that this estimator can be evaluated at a single point by finding the solution to a linear program. We then show that, under some mild regularity conditions, our estimator can be used to consistently estimate both the convex function and its subdifferentials.

Our work goes beyond those mentioned above in the following ways: our method does not require any tuning parameter(s), which is a major drawback for most nonparametric regression methods, such as kernel-based procedures. The choice of the tuning parameter(s) is especially problematic in higher dimensions; for example, kernel based methods would require the choice of a $d \times d$ matrix of bandwidths. The sets of assumptions that most authors have used to study the estimation of a multidimensional convex regression function are more restrictive and of a different nature than the ones in this paper. As opposed to the maximum likelihood approach used in Banker and Maindiratta (1992), Matzkin (1991), Allon et al. (2007) and Sarath and Maindiratta (1997), we prove the consistency of the estimator keeping the distribution of the errors completely *unspecified*; for example, in the i.i.d. case we only assume that the errors have zero expectation and finite second moment. The estimators in Beresteanu (2007) are sieved least squares estimators and assume that the observed values of the predictors lie on equidistant grids of rectangular domains. By contrast, our estimators are unsieved and our assumptions on the spatial arrangement of the predictor values are much more relaxed. In fact, we prove the consistency of the least squares estimator under both fixed and stochastic design settings; we also allow for heteroscedastic errors. In addition, we show that the least squares estimator can also be used to approximate the gradients and subdifferentials of the underlying convex function.

It is hard to overstate the importance of convex functions in applied mathematics. For instance, optimization problems with convex objective functions over convex sets appear in many applications. Thus, the question of accurately estimating a convex regression function is indeed interesting from a theoretical perspective. However, it turns out that convex regression is important for numerous reasons besides statistical curiosity. Convexity also appears in many applied sciences. One such field of application is microeconomic theory. Production functions are often supposed to be concave and componentwise nondecreasing. In this context, concavity reflects decreasing marginal returns. Concavity also plays a role in the theory of rational choice since it is a common assumption for utility functions, on which it represents decreasing marginal utility. The interested reader can see Hildreth (1954), Varian (1982) or Varian (1984) for more information regarding the importance of concavity/convexity in economic theory.

The paper is organized as follows. In Section 2 we discuss the estimation procedure, characterize the estimator and show how it can be computed by solving a positive semidefinite quadratic program and a linear program. Section 3 starts with a description of the deterministic and stochastic design regression schemes. The statements and proofs of our main results are also included in Section 3. In Section 4 we provide the proofs of some technical lemmas used to prove the main theorem. Although we have omitted the proofs of some auxiliary results, they can be found in the supplemental document [Seijo and Sen (2010)].

2. Characterization and finite sample properties. We start with some notation. For convenience, we will regard elements of the Euclidean space $\mathbb{R}^m$ as column vectors and denote their components with upper indices, that is, any $z \in \mathbb{R}^m$ will be denoted by $z = (z^1, z^2, \ldots, z^m)'$. The symbol $\overline{\mathbb{R}}$ will stand for the extended real line. Additionally, for any set $A \subset \mathbb{R}^d$ we will denoted as $\mathrm{Conv}(A)$ its convex hull and we will write $\mathrm{Conv}(X_1, \ldots, X_n)$ instead of $\mathrm{Conv}(\{X_1, \ldots, X_n\})$. Finally, we will use $\langle \cdot, \cdot \rangle$ and $|\cdot|$ to denote the standard inner product and norm in Euclidean spaces, respectively.

For $\mathcal{X} = \{X_1, \ldots, X_n\} \subset \mathfrak{X} \subset \mathbb{R}^d$, consider the set $\mathcal{K}_{\mathcal{X}}$ of all vectors $z = (z^1, \ldots, z^n)' \in \mathbb{R}^n$ for which there is a convex function $\psi : \mathfrak{X} \to \mathbb{R}$ such that $\psi(X_j) = z^j$ for all $j = 1, \ldots, n$. Then, a necessary and sufficient condition for a convex function ψ to minimize the sum of squared errors is that $\psi(X_j) = Z_n^j$ for $j = 1, \ldots, n$, where

$$Z_n = \operatorname*{arg\,min}_{z \in \mathcal{K}_{\mathcal{X}}} \left\{ \sum_{k=1}^{n} |Y_k - z^k|^2 \right\}. \tag{2}$$

The computation of the vector Z_n is crucial for the estimation procedure. We will show that such a vector exists and is unique. However, it should be noted that there are many convex functions ψ satisfying $\psi(X_j) = Z_n^j$ for all $j = 1, \ldots, n$. Although any of these functions can play the role of the least squares estimator, there is one such function which is easily evaluated in $\mathrm{Conv}(X_1, \ldots, X_n)$. For computational convenience, we will define our least squares estimator $\hat{\phi}_n$ to be precisely this function and describe it explicitly in (7) and the subsequent discussion.

In what follows we show that both the vector Z_n and the least squares estimator $\hat{\phi}_n$ are well defined for any n data points $(X_1, Y_1), \ldots, (X_n, Y_n)$. We will also provide two characterizations of the set $\mathcal{K}_{\mathcal{X}}$ and show that the vector Z_n can be computed by solving a positive semidefinite quadratic program. Finally, we will prove that for any $x \in \mathrm{Conv}(X_1, \ldots, X_n)$ one can obtain $\hat{\phi}_n(x)$ by solving a linear program.

2.1. *Existence and uniqueness.* We start with two characterizations of the set $\mathcal{K}_{\mathcal{X}}$. The developments here are similar to those in Allon et al. (2007) and Kuosmanen (2008).

LEMMA 2.1 (Primal characterization). *Let $z \in \mathbb{R}^n$. Then, $z \in \mathcal{K}_{\mathcal{X}}$ if and only if for every $j = 1, \ldots, n$, the following holds*:

$$z^j = \inf\left\{\sum_{k=1}^n \theta^k z^k : \sum_{k=1}^n \theta^k = 1, \sum_{k=1}^n \theta^k X_k = X_j, \theta \geq 0, \theta \in \mathbb{R}^n\right\}, \tag{3}$$

where the inequality $\theta \geq 0$ holds componentwise.

PROOF. Define the function $g : \mathbb{R}^d \to \overline{\mathbb{R}}$ by

$$g(x) = \inf\left\{\sum_{k=1}^n \theta^k z^k : \sum_{k=1}^n \theta^k = 1, \sum_{k=1}^n \theta^k X_k = x, \theta \geq 0, \theta \in \mathbb{R}^n\right\}, \tag{4}$$

where we use the convention that $\inf(\varnothing) = +\infty$. By Lemma 2.1 in Seijo and Sen (2010), g is convex and finite on the X_j's. Hence, if z^j satisfies (3), then $z^j = g(X_j)$ for every $j = 1, \ldots, n$ and it follows that $z \in \mathcal{K}_{\mathcal{X}}$.

Conversely, assume that $z \in \mathcal{K}_{\mathcal{X}}$ and $g(X_j) \neq z^j$ for some j. Note that $g(X_k) \leq z^k$ for any k from the definition of g. Thus, we may suppose that $g(X_j) < z^j$. As $z \in \mathcal{K}_{\mathcal{X}}$, there is a convex function ψ such that $\psi(X_k) = z^k$ for all $k = 1, \ldots, n$. Then, from the definition of $g(X_j)$ there exist $\theta_0 \in \mathbb{R}^n$ with $\theta_0 \geq 0$ and $\theta_0^1 + \cdots + \theta_0^n = 1$ such that $\theta_0^1 X_1 + \cdots + \theta_0^n X_n = X_j$ and

$$\sum_{k=1}^n \theta_0^k \psi(X_k) = \sum_{k=1}^n \theta_0^k z^k < z^j = \psi(X_j) = \psi\left(\sum_{k=1}^n \theta_0^k X_k\right),$$

which leads to a contradiction because ψ is convex. □

We now provide an alternative characterization of the set $\mathcal{K}_{\mathcal{X}}$ based on the dual problem to the linear program used in Lemma 2.1.

LEMMA 2.2 (Dual characterization). *Let $z \in \mathbb{R}^n$. Then, $z \in \mathcal{K}_{\mathcal{X}}$ if and only if for any $j = 1, \ldots, n$ we have*

$$z^j = \sup\{\langle \xi, X_j \rangle + \eta : \langle \xi, X_k \rangle + \eta \leq z^k \ \forall k = 1, \ldots, n, \xi \in \mathbb{R}^d, \eta \in \mathbb{R}\}. \tag{5}$$

Moreover, $z \in \mathcal{K}_{\mathcal{X}}$ if and only if there exist vectors $\xi_1, \ldots, \xi_n \in \mathbb{R}^d$ such that

$$\langle \xi_j, X_k - X_j \rangle \leq z^k - z^j \qquad \forall k, j \in \{1, \ldots, n\}. \tag{6}$$

PROOF. According to the primal characterization, $z \in \mathcal{K}_{\mathcal{X}}$ if and only if the linear programs defined by (3) have the z^j's as optimal values. The linear programs in (5) are the dual problems to those in (3). Then, the duality theorem for linear programs [see Luenberger (1984), page 89] implies that $z \in \mathcal{K}_{\mathcal{X}}$ if and only if the z^j's are the optimal values to the programs in (5).

To prove the second assertion let us first assume that $z \in \mathcal{K}_{\mathcal{X}}$. For each $j \in \{1, \ldots, n\}$ take any solution (ξ_j, η_j) to (5). Then by (5), $\eta_j = z^j - \langle \xi_j, X_j \rangle$ and the

inequalities in (6) follow immediately because we must have $\langle \xi_j, X_k \rangle + \eta_j \leq z^k$ for any $k \in \{1, \ldots, n\}$. Conversely, take $z \in \mathbb{R}^n$ and assume that there are $\xi_1, \ldots, \xi_n \in \mathbb{R}^d$ satisfying (6). Take any $j \in \{1, \ldots, n\}$, $\eta_j = z^j - \langle \xi_j, X_j \rangle$ and θ to be the vector in $\mathbb{R}^n$ with components $\theta^k = \delta_{kj}$, where δ_{kj} is the Kronecker δ. It follows that $\langle \xi_j, X_k \rangle + \eta_j \leq z^k \ \forall k = 1, \ldots, n$ so (ξ_j, η_j) is feasible for the linear program in (5). In addition, θ is feasible for the linear program in (3) so the weak duality principle of linear programming [see Luenberger (1984), Lemma 1, page 89] implies that $\langle \xi, X_j \rangle + \eta \leq z^j$ for any pair (ξ, η) which is feasible for the problem in the right-hand side of (5). We thus have that z^j is an upper bound attained by the feasible pair (ξ_j, η_j) and hence (5) holds for all $j = 1, \ldots, n$. $\square$

Both the primal and dual characterizations are useful for our purposes. The primal plays a key role in proving the existence and uniqueness of the least squares estimator. The dual is crucial for its computation.

LEMMA 2.3. *The set $\mathcal{K}_{\mathcal{X}}$ is a closed, convex cone in $\mathbb{R}^n$ and the vector Z_n satisfying (2) is uniquely defined.*

PROOF. That $\mathcal{K}_{\mathcal{X}}$ is a convex cone follows trivially from the definition of the set. Now, if $z \notin \mathcal{K}_{\mathcal{X}}$, then there is $j \in \{1, \ldots, n\}$ for which $z^j > g(X_j)$ with the function g defined as in (4). Thus, there is $\theta_0 \in \mathbb{R}^n$ with $\theta_0 \geq 0$ and $\theta_0^1 + \cdots + \theta_0^n = 1$ such that $\theta_0^1 X_1 + \cdots + \theta_0^n X_n = X_j$ and $\sum_{k=1}^n \theta_0^k z^k < z^j$. Setting $\delta = \frac{1}{2}(z^j - \sum_{k=1}^n \theta_0^k z^k)$ it is easily seen that for all $\zeta \in \prod_{k=1}^n (z^k - \delta, z^k + \delta)$ we still have $\sum_{k=1}^n \theta_0^k \zeta^k < \zeta^j$ and thus $\zeta \notin \mathcal{K}_{\mathcal{X}}$. Thus, we have shown that for any $z \notin \mathcal{K}_{\mathcal{X}}$ there is a neighborhood U of z with $U \subset \mathbb{R}^n \setminus \mathcal{K}_{\mathcal{X}}$. Therefore, $\mathcal{K}_{\mathcal{X}}$ is closed and the vector Z_n is uniquely determined as the projection of $(Y_1, \ldots, Y_n) \in \mathbb{R}^n$ onto the closed convex set $\mathcal{K}_{\mathcal{X}}$ [see Conway (1985), Theorem 2.5, page 9]. $\square$

We are now in a position to define the least squares estimator. Given observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ from model (1), we take the nonparametric least squares estimator to be the function $\hat{\phi}_n : \mathbb{R}^d \to \mathbb{R}$ defined by

$$\hat{\phi}_n(x) = \inf \left\{ \sum_{k=1}^n \theta^k Z_n^k : \sum_{k=1}^n \theta^k = 1, \sum_{k=1}^n \theta^k X_k = x, \theta \geq 0, \theta \in \mathbb{R}^n \right\} \tag{7}$$

for any $x \in \mathbb{R}^d$. Here we are taking the convention that $\inf(\varnothing) = +\infty$. This function is well defined because the vector Z_n exists and is unique for the sample. The estimator is, in fact, a polyhedral convex function [i.e., a convex function whose epigraph is a polyhedral; see Rockafellar (1970), page 172] and satisfies, as a consequence of Lemma 2.1 in Seijo and Sen (2010),

$$\hat{\phi}_n(x) = \sup_{\psi \in \mathcal{K}_{\mathcal{X}, Z_n}} \{\psi(x)\},$$

where $\mathcal{K}_{\mathcal{X},Z_n}$ is the collection of all convex functions $\psi : \mathbb{R}^d \to \mathbb{R}$ such that $\psi(X_j) \le Z_n^j$ for all $j = 1, \ldots, n$. Thus, $\hat{\phi}_n$ is the largest convex function that never exceeds the Z_n^j's. It is immediate that $\hat{\phi}_n$ is indeed a convex function (as the supremum of any family of convex functions is itself convex). The primal characterization of the set $\mathcal{K}_{\mathcal{X}}$ implies that $\hat{\phi}_n(X_j) = Z_n^j$ for all $j = 1, \ldots, n$.

2.2. *Finite sample properties.* In the following lemma we state some of the most important finite sample properties of the least squares estimator defined in (7). For a proof see Lemma 2.2 of Seijo and Sen (2010).

LEMMA 2.4. *Let $\hat{\phi}_n$ be the least squares estimator obtained from the sample $(X_1, Y_1), \ldots, (X_n, Y_n)$. Then*:

(i) $\sum_{k=1}^n (\psi(X_k) - \hat{\phi}_n(X_k))(Y_k - \hat{\phi}_n(X_k)) \le 0$ *for any convex function ψ which is finite on* $\mathrm{Conv}(X_1, \ldots, X_n)$;

(ii) $\sum_{k=1}^n \hat{\phi}_n(X_k)(Y_k - \hat{\phi}_n(X_k)) = 0$;

(iii) $\sum_{k=1}^n Y_k = \sum_{k=1}^n \hat{\phi}_n(X_k)$;

(iv) *the set on which $\hat{\phi}_n < \infty$ is* $\mathrm{Conv}(X_1, \ldots, X_n)$;

(v) *for any $x \in \mathbb{R}^d$ the map $(X_1, \ldots, X_n, Y_1, \ldots, Y_n) \hookrightarrow \hat{\phi}_n(x)$ is a Borel-measurable function from $\mathbb{R}^{n(d+1)}$ into $\mathbb{R}$.*

2.3. *Computation of the estimator.* Once the vector Z_n defined in (2) has been obtained, the evaluation of $\hat{\phi}_n$ at a single point x can be carried out by solving the linear program in (7). Thus, we need to find a way to compute Z_n. And here the dual characterization proves of vital importance, since it allows us to compute Z_n by solving a quadratic program.

LEMMA 2.5. *Consider the positive semidefinite quadratic program*

$$\min \sum_{k=1}^n |Y_k - z^k|^2 \qquad \text{subject to } \langle \xi_k, X_j - X_k \rangle \le z^j - z^k \quad \forall k, j = 1, \ldots, n, \xi_1, \ldots, \xi_n \in \mathbb{R}^d, z \in \mathbb{R}^n. \tag{8}$$

Then, this program has a unique solution Z_n in z, that is, for any two solutions $(\xi_1, \ldots, \xi_n, z)$ and $(\tau_1, \ldots, \tau_n, \zeta)$ we have $z = \zeta = Z_n$. This solution Z_n is the only vector in $\mathbb{R}^n$ which satisfies (2).

PROOF. From Lemma 2.2, if $(\xi_1, \ldots, \xi_n, z)$ belongs in the feasible set of this program, then $z \in \mathcal{K}_{\mathcal{X}}$. Moreover, for any $z \in \mathcal{K}_{\mathcal{X}}$ there are $\xi_1, \ldots, \xi_n \in \mathbb{R}^d$ such that $(\xi_1, \ldots, \xi_n, z)$ belongs to the feasible set of the quadratic program. Since the objective function only depends on z, solving the quadratic program is the same as getting the element of $\mathcal{K}_{\mathcal{X}}$ which is the closest to Y. This element is, of course, the uniquely defined Z_n satisfying (2). □

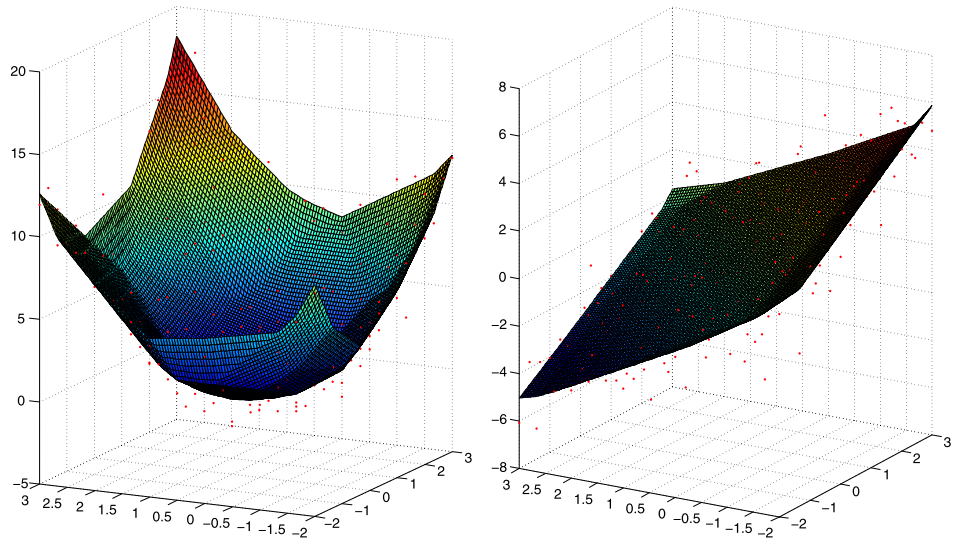

FIG. 1. *The scatterplot and nonparametric least squares estimator of the convex regression function when* (a) $\phi(x) = |x|^2$ (*left panel*); (b) $\phi(x) = -x^1 + x^2$ (*right panel*).

The quadratic program (8) is positive semidefinite. This implies certain computational complexities, but most modern nonlinear programming solvers can handle this type of optimization problems. Some examples of high-performance quadratic programming solvers are CPLEX, LINDO, MOSEK and QPOPT. Here we present two simulated examples to illustrate the computation of the estimator when $d = 2$. The first one, depicted in Figure 1(a), corresponds to the case where $\phi(x) = |x|^2$. Figure 1(b) shows the convex function estimator when the regression function is the hyperplane $\phi(x) = -x^1 + x^2$. In both cases, $n = 256$ observations were used and the errors were assumed to be i.i.d. from the standard normal distribution. All the computations were carried out using the MOSEK optimization toolbox for Matlab and the run time for each example was less than 2 minutes on a standard desktop PC. We refer the reader to Kuosmanen (2008) for additional numerical examples (although the examples there are for the estimation of concave, componentwise nondecreasing functions, the computational complexities are the same).

3. Consistency of the least squares estimator. The main goal of this paper is to show that in an appropriate setting the nonparametric least squares estimator $\hat{\phi}_n$ described above is consistent for estimating the convex function ϕ on the set $\mathfrak{X}$. In this context, we will prove the consistency of $\hat{\phi}_n$ in both fixed and stochastic design regression settings.

Before proceeding any further we would like to introduce some notation. For any Borel set $\mathfrak{X} \subset \mathbb{R}^d$ we will denote by $\mathcal{B}_{\mathfrak{X}}$ the σ-algebra of Borel subsets of $\mathfrak{X}$. Given a sequence of events $(A_n)_{n=1}^{\infty}$ we will be using the notation $[A_n \text{ i.o.}]$ and $[A_n \text{ a.a.}]$ to denote $\overline{\lim}\, A_n$ and $\underline{\lim}\, A_n$, respectively.

Now, consider a convex function $f:\mathbb{R}^d \to \overline{\mathbb{R}}$. This function is said to be proper if $f(x) > -\infty$ for every $x \in \mathbb{R}^d$. The effective domain of f, denoted by $\mathrm{Dom}(f)$, is the set of points $x \in \mathbb{R}^d$ for which $f(x) < \infty$. The subdifferential of f at a point $x \in \mathbb{R}^d$ is the set $\partial f(x) \subset \mathbb{R}^d$ of all vectors ξ satisfying the inequalities

$$\langle \xi, h\rangle \le f(x+h) - f(x) \qquad \forall h \in \mathbb{R}^d.$$

The elements of $\partial f(x)$ are called subgradients of f at x [see Rockafellar (1970)]. For a set $A \subset \mathbb{R}^d$ we denote by A°, $\overline{A}$ and ∂A its interior, closure and boundary, respectively. We write $\mathrm{Ext}(A) = \mathbb{R}^d \setminus \overline{A}$ for the exterior of the set A and $\mathrm{diam}(A) := \sup_{x,y\in A} |x-y|$ for the diameter of A. We also use the sup-norm notation, that is, for a function $g:\mathbb{R}^d \to \mathbb{R}$ we write $\|g\|_A = \sup_{x\in A} |g(x)|$.

To avoid measurability issues regarding some sets, especially those involving the random set-valued functions $\{\partial\hat{\phi}_n(x)\}_{x\in\mathfrak{X}^\circ}$, we will use the symbols $\mathbf{P}_*$ and $\mathbf{P}^*$ to denote inner and outer probabilities, respectively. We refer the reader to Van der Vaart and Wellner (1996), pages 6–15, for the basic properties of inner and outer probabilities. In this context, a sequence of (not necessarily measurable) functions $(\Psi_n)_{n=1}^\infty$ from a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ into $\mathbb{R}$ is said to converge to a function Ψ *almost surely* [see Van der Vaart and Wellner (1996), Definition 1.9.1(iv), page 52], written $\Psi_n \xrightarrow{\text{a.s.}} \Psi$, if $\mathbf{P}_*(\Psi_n \to \Psi) = 1$. We will use the standard notation $\mathbf{P}(A)$ for the probabilities of all events A whose measurability can be easily inferred from the measurability of the random variables $\{\hat{\phi}_n(x)\}_{x\in\mathfrak{X}}$, established in Lemma 2.4.

Our main theorems hold for both fixed and stochastic design schemes, and the proofs are very similar. They differ only in minor steps. Therefore, for the sake of simplicity, we will denote the observed values of the regressor variables always with the capital letters X_n. For any Borel set $\mathtt{X} \subset \mathbb{R}^d$, we write

$$N_n(\mathtt{X}) = \#\{1 \le j \le n : X_j \in \mathtt{X}\}.$$

The quantities X_n and $N_n(\mathtt{X})$ are nonrandom under the fixed design but random under the stochastic one.

3.1. *Fixed design.* In a "fixed design" regression setting we assume that the regressor values are nonrandom and that all the uncertainty in the model comes from the response variable. We will now list a set of assumptions for this type of design. The one-dimensional case has been proven, under different regularity conditions, in Hanson and Pledger (1976).

(A1) We assume that we have a sequence $(X_n, Y_n)_{n=1}^\infty$ satisfying

$$Y_k = \phi(X_k) + \varepsilon_k,$$

where $(\varepsilon_n)_{n=1}^\infty$ is an i.i.d. sequence with $\mathbf{E}(\varepsilon_j) = 0$, $\mathbf{E}(\varepsilon_j^2) = \sigma^2 < \infty$ and $\phi:\mathbb{R}^d \to \mathbb{R}$ is a proper convex function.

(A2) The nonrandom sequence $(X_n)_{n=1}^{\infty}$ is contained in a closed, convex set $\mathfrak{X} \subset \mathbb{R}^d$ with $\mathfrak{X}^\circ \neq \varnothing$ and $\mathfrak{X} \subset \mathrm{Dom}(\phi)$.

(A3) We assume the existence of a Borel measure ν on $\mathfrak{X}$ satisfying:
(i) $\{\mathtt{X} \in \mathcal{B}_{\mathfrak{X}} : \nu(\mathtt{X}) = 0\} = \{\mathtt{X} \in \mathcal{B}_{\mathfrak{X}} : \mathtt{X} \text{ has Lebesgue measure } 0\}$.
(ii) $\frac{1}{n} N_n(\mathtt{X}) \to \nu(\mathtt{X})$ for any Borel set $\mathtt{X} \subset \mathfrak{X}$.

Condition (A1) may be replaced by the following:

(A4) We assume that we have a sequence $(X_n, Y_n)_{n=1}^{\infty}$ satisfying

$$Y_k = \phi(X_k) + \varepsilon_k,$$

where $\phi : \mathbb{R}^d \to \mathbb{R}$ is a proper convex function and $(\varepsilon_n)_{n=1}^{\infty}$ is an independent sequence of random variables satisfying:
(i) $\mathbf{E}(\varepsilon_n) = 0 \ \forall n \in \mathbb{N}$ and $\underline{\lim} \frac{1}{n} \sum_{k=1}^{n} \mathbf{E}(|\varepsilon_k|) > 0$.
(ii) $\sum_{n=1}^{\infty} \frac{\mathrm{Var}(\varepsilon_n^2)}{n^2} < \infty$.
(iii) $\sup_{n \in \mathbb{N}} \{\mathbf{E}(\varepsilon_n^2)\} < \infty$.
Under these conditions we define $\sigma^2 := \overline{\lim}_{n \to \infty} \frac{1}{n} \sum_{j=1}^{n} \mathbf{E}(\varepsilon_j^2)$.

The raison d'etre of condition (A4) is to allow the variance of the error terms to depend on the regressors. We make the distinction between (A1) and (A4) because in the case of i.i.d. errors it suffices to require a finite second moment to ensure consistency.

3.2. *Stochastic design.* In this setting we assume that $(X_n, Y_n)_{n=1}^{\infty}$ is an i.i.d. sequence from some Borel probability measure μ on $\mathbb{R}^{d+1}$. Here we make the following assumptions on the measure μ:

(A5) There is a closed, convex set $\mathfrak{X} \subset \mathbb{R}^d$ with $\mathfrak{X}^\circ \neq \varnothing$ such that $\mu(\mathfrak{X} \times \mathbb{R}) = 1$. Also,

$$\int_{\mathfrak{X} \times \mathbb{R}} y^2 \mu(dx, dy) < \infty.$$

(A6) There is a proper convex function $\phi : \mathbb{R}^d \to \mathbb{R}$ with $\mathfrak{X} \subset \mathrm{Dom}(\phi)$ such that whenever $(X, Y) \sim \mu$ we have $\mathbf{E}(Y - \phi(X)|X) = 0$ and $\mathbf{E}(|Y - \phi(X)|^2) = \sigma^2 < \infty$. Thus, ϕ is the regression function.

(A7) Denoting by $\nu(\cdot) := \mu((\cdot) \times \mathbb{R})$ the x-marginal of μ, we assume that

$$\{\mathtt{X} \in \mathcal{B}_{\mathfrak{X}} : \nu(\mathtt{X}) = 0\} = \{\mathtt{X} \in \mathcal{B}_{\mathfrak{X}} : \mathtt{X} \text{ has Lebesgue measure } 0\}.$$

Observe that conditions (A5)–(A7) allow for stochastic dependency between the error variable $Y - \phi(X)$ and the regressor X. Although some level of dependency can be put to satisfy conditions (A2)–(A4), the measure μ allows us to take into account some cases which would not fit in the fixed design setting (even by conditioning on the regressors).

3.3. *Main results.* We can now state the two main results of this paper. The first result shows that assuming only the convexity of ϕ, the least squares estimator can be used to consistently estimate both ϕ and its subdifferentials $\partial\phi(x)$.

THEOREM 3.1. *Under any of* (A1)–(A3), (A2)–(A4) *or* (A5)–(A7) *we have*:

(i) $\mathbf{P}(\sup_{x\in \mathrm{X}}\{|\hat{\phi}_n(x) - \phi(x)|\} \to 0$ *for any compact set* $\mathrm{X} \subset \mathfrak{X}^\circ) = 1$.

(ii) *For every* $x \in \mathfrak{X}^\circ$ *and every* $\xi \in \mathbb{R}^d$

$$\overline{\lim_{n\to\infty}} \lim_{h\downarrow 0} \frac{\hat{\phi}_n(x+h\xi) - \hat{\phi}_n(x)}{h} \le \lim_{h\downarrow 0} \frac{\phi(x+h\xi) - \phi(x)}{h} \qquad \textit{almost surely}.$$

(iii) *Denoting by* $\boldsymbol{B}$ *the unit ball* (*w.r.t. the Euclidean norm*) *we have*

$$\mathbf{P}_*\big(\partial\hat{\phi}_n(x) \subset \partial\phi(x) + \varepsilon\mathbf{B} \ a.a.\big) = 1 \qquad \forall\varepsilon > 0, \forall x \in \mathfrak{X}^\circ.$$

(iv) *If* ϕ *is differentiable at* $x \in \mathfrak{X}^\circ$, *then*

$$\sup_{\xi\in\partial\hat{\phi}_n(x)} \{|\xi - \nabla\phi(x)|\} \xrightarrow{a.s.} 0.$$

Our second result states that assuming differentiability of ϕ on the entire $\mathfrak{X}^\circ$ allows us to use the subdifferentials of the least squares estimator to consistently estimate $\nabla\phi$ uniformly on compact subsets of $\mathfrak{X}^\circ$.

THEOREM 3.2. *If* ϕ *is differentiable on* $\mathfrak{X}^\circ$, *then under any of* (A1)–(A3), (A2)–(A4) *or* (A5)–(A7) *we have*

$$\mathbf{P}_*\Big(\sup_{\substack{\xi\in\partial\hat{\phi}_n(x)\\ x\in\mathrm{X}}} \{|\xi - \nabla\phi(x)|\} \to 0 \textit{ for any compact set } \mathrm{X} \subset \mathfrak{X}^\circ\Big) = 1.$$

3.4. *Proof of the main results.* Before embarking on the proofs, one must notice that there are some statements which hold true under any of (A1)–(A3), (A2)–(A4) or (A5)–(A7). We list the most important ones below, since they will be used later.

- For any set $\mathrm{X} \subset \mathfrak{X}$ we have

$$\frac{N_n(\mathrm{X})}{n} \xrightarrow{\text{a.s.}} \nu(\mathrm{X}). \tag{9}$$

- The strong law of large numbers implies that for any Borel set $\mathrm{X} \subset \mathfrak{X}$ with positive Lebesgue measure we have

$$\frac{1}{N_n(\mathrm{X})} \sum_{\substack{X_k\in\mathrm{X}\\ 1\le k\le n}} \big(Y_k - \phi(X_k)\big) \xrightarrow{\text{a.s.}} 0 \tag{10}$$

and also

$$\overline{\lim_{n\to\infty}} \frac{1}{n} \sum_{1\le k\le n} \big(Y_k - \phi(X_k)\big)^2 = \sigma^2 \qquad \text{a.s.} \tag{11}$$

We would like to point out that in the case of condition (A4), (A4)(iii) allows us to obtain (10) from an application of a version of the strong law of large number for uncorrelated random variables, as it appears in Chung (2001), page 108, Theorem 5.1.2. Similarly, condition (A4)(ii) implies that we can apply a version the strong law of large numbers for independent random variables as in Williams (1991), Lemma 12.8, page 118, or in Folland (1999), Theorem 10.12, page 322, to obtain (11).

- For any Borel subset $\mathrm{X} \subset \mathfrak{X}$ with positive Lebesgue measure,

$$\#\{n \in \mathbb{N} : X_n \in \mathrm{X}\} \xrightarrow{\text{a.s.}} +\infty. \tag{12}$$

PROOF OF THEOREM 3.1. We will only make distinctions among the design schemes in the proof when we use any property besides (9), (10), (11) or (12). For the sake of clarity, we divide the proof in steps.

Step I: We start by showing that for any set with positive Lebesgue measure there is a uniform band around the regression function (over that set) such that $\hat{\phi}_n$ comes within the band at least at one point for all but finitely many n's. This fact is stated in the following lemma (proved in Section 4.1).

LEMMA 3.1. *For any set $\mathrm{X} \subset \mathfrak{X}$ with positive Lebesgue measure we have*

$$\mathbf{P}\Big(\inf_{x\in\mathrm{X}}\{|\hat{\phi}_n(x) - \phi(x)|\} \ge M \text{ i.o.}\Big) = 0 \qquad \forall M > \frac{\sigma}{\sqrt{\nu(\mathrm{X})}}.$$

Step II: The idea is now to use the convexity of both ϕ and $\hat{\phi}_n$, to show that the previous result in fact implies that the sup-norm of $\hat{\phi}_n$ is uniformly bounded on compact subsets of $\mathfrak{X}^\circ$. We achieve this goal in the following two lemmas [whose proofs are given in Sections 1.1 and 1.2 of Seijo and Sen (2010), resp.].

LEMMA 3.2. *Let $\mathrm{X} \subset \mathfrak{X}^\circ$ be compact with positive Lebesgue measure. Then, there is a positive real number K_{X} such that*

$$\mathbf{P}\Big(\inf_{x\in\mathrm{X}}\{\hat{\phi}_n(x)\} \le -K_{\mathrm{X}} \text{ i.o.}\Big) = 0.$$

LEMMA 3.3. *Let $\mathrm{X} \subset \mathfrak{X}^\circ$ be a compact set with positive Lebesgue measure. Then, there is $K_{\mathrm{X}} > 0$ such that*

$$\mathbf{P}\Big(\sup_{x\in\mathrm{X}}\{\hat{\phi}_n(x)\} \ge K_{\mathrm{X}} \text{ i.o.}\Big) = 0.$$

Step III: Convex functions are determined by their subdifferential mappings [see Rockafellar (1970), Theorem 24.9, page 239]. Moreover, having a uniform upper bound $K_{\mathtt{X}}$ for the norms of all the subgradients over a compact region $\mathtt{X}$ imposes a Lipschitz continuity condition on the convex function over $\mathtt{X}$ [see Rockafellar (1970), Theorem 24.7, page 237], the Lipschitz constant being $K_{\mathtt{X}}$. For these reasons, it is important to have a uniform upper bound on the norms of the subgradients of $\hat{\phi}_n$ on compact regions. The following lemma [proved in Section 1.3 of Seijo and Sen (2010)] states that this can be achieved.

LEMMA 3.4. *Let $\mathtt{X} \subset \mathfrak{X}^\circ$ be a compact set with positive Lebesgue measure. Then, there is $K_{\mathtt{X}} > 0$ such that*

$$\mathbf{P}^*\Big(\sup_{\substack{\xi \in \partial \hat{\phi}_n(x) \\ x \in \mathtt{X}}} \{|\xi|\} > K_{\mathtt{X}} \ i.o. \Big) = 0.$$

Step IV: For the next results we need to introduce some further notation. We will denote by μ_n the empirical measure defined on $\mathbb{R}^{d+1}$ by the sample $(X_1, Y_1), \ldots, (X_n, Y_n)$. In agreement with Van der Vaart and Wellner (1996), Definition 2.1.5, page 83, given a class of functions $\mathcal{G}$ on $D \subset \mathbb{R}^{d+1}$, a seminorm $\|\cdot\|$ on some space containing $\mathcal{G}$ and $\varepsilon > 0$ we denote by $N(\varepsilon, \mathcal{G}, \|\cdot\|)$ the ε-covering number of $\mathcal{G}$ with respect to $\|\cdot\|$.

Although Lemmas 3.5 and 3.7 may seem unrelated to what has been done so far, they are crucial for the further developments. Lemma 3.5 (proved in Section 4.2) shows that the class of convex functions is not very complex in terms of entropy. Lemma 3.7 is a uniform version of the strong law of large numbers which proves vital in the proof of Lemma 3.8.

LEMMA 3.5. *Let $\mathtt{X} \subset \mathfrak{X}^\circ$ be a compact rectangle with positive Lebesgue measure. For $K > 0$ consider the class $\mathcal{G}_{K,\mathtt{X}}$ of all functions of the form $\psi(X)(Y - \phi(X))\mathbf{1}_{\mathtt{X}}(X)$ where ψ ranges over the class $\mathcal{D}_{K,\mathtt{X}}$ of all proper convex functions which satisfy*:

(a) $\|\psi\|_{\mathtt{X}} \le K$;
(b) $\bigcup_{\xi \in \partial \psi(x), x \in \mathtt{X}} \{\xi\} \subset [-K, K]^d$.

Then, for any $\varepsilon > 0$ we have

$$\overline{\lim_{n \to \infty}} \, N\big(\varepsilon, \mathcal{G}_{K,\mathtt{X}}, \mathbb{L}_1(\mathtt{X} \times \mathbb{R}, \mu_n)\big) < \infty \qquad \textit{almost surely},$$

and there is a positive constant $A_\varepsilon < \infty$, depending only on $(X_1, \ldots, X_n)$, K and $\mathtt{X}$, such that the covering numbers $N(\frac{\varepsilon}{n} \sum_{j=1}^n |Y_j - \phi(X_j)|, \mathcal{G}_{K,\mathtt{X}}, \mathbb{L}_1(\mathtt{X} \times \mathbb{R}, \mu_n))$ are bounded above by A_ε, for all $n \in \mathbb{N}$, almost surely.

The proofs of Lemmas 3.7 and 3.8 (given in Sections 4.4 and 4.5, resp.) are the only parts in the whole proof where we must treat the different design schemes

separately. To make the argument work, a small lemma (proved in Section 4.3) for the set of conditions (A2)–(A4) is required. We include it here for the sake of completeness and to point out the difference between the schemes.

LEMMA 3.6. *Consider the set of conditions* (A2)–(A4) *and a subsequence* $(n_k)_{k=1}^{\infty}$ *such that*

$$\lim_{k\to\infty} \frac{1}{n_k} \sum_{j=1}^{n_k} \mathbf{E}(\varepsilon_j^2) = \sigma^2.$$

Let $(\mathrm{X}_m)_{m=1}^{\infty}$ *be an increasing sequence of compact subsets of* $\mathfrak{X}$ *satisfying* $\nu(\mathrm{X}_m) \to 1$. *Then*,

$$\lim_{m\to\infty} \underline{\lim}_{k\to\infty} \frac{1}{n_k} \sum_{\{1\le j\le n_k\,:\,X_j\in \mathrm{X}_m\}} \mathbf{E}(\varepsilon_j^2) = \sigma^2.$$

We are now ready to state the key result on the uniform law of large numbers. We refer the reader to Section 4.4 for a complete proof.

LEMMA 3.7. *Consider the notation of Lemma* 3.5 *and let* $\mathrm{X} \subset \mathfrak{X}^{\circ}$ *be any finite union of compact rectangles with positive Lebesgue measure. Then*,

$$\sup_{\psi\in\mathcal{D}_{K,\mathrm{X}}} \left\{ \left| \frac{1}{n} \sum_{\{1\le j\le n\,:\,X_j\in\mathrm{X}\}} \psi(X_j)\big(Y_j - \phi(X_j)\big) \right| \right\} \xrightarrow{a.s.} 0.$$

Step V: With the aid of all the results proved up to this point, it is now possible to show that Lemma 3.1 is in fact true if we replace M by an arbitrarily small $\eta > 0$. The proof of the following lemma is given in Section 4.5.

LEMMA 3.8. *Let* $\mathrm{X} \subset \mathfrak{X}^{\circ}$ *be any compact set with positive Lebesgue measure. Then*:

(i) $\mathbf{P}(\inf_{x\in\mathrm{X}}\{\phi(x) - \hat{\phi}_n(x)\} \ge \eta \ i.o.) = 0 \ \forall \eta > 0$,
(ii) $\mathbf{P}(\sup_{x\in\mathrm{X}}\{\phi(x) - \hat{\phi}_n(x)\} \le -\eta \ i.o.) = 0 \ \forall \eta > 0$.

Step VI: Combining the last lemma with the fact that we have a uniform bound on the norms of the subgradients on compacts, we can state and prove the consistency result on compacts. This is done in the next lemma (proof included in Section 4.6).

LEMMA 3.9. *Let* $\mathrm{X} \subset \mathfrak{X}^{\circ}$ *be a compact set with positive Lebesgue measure. Then*:

(i) $\mathbf{P}(\inf_{x\in\mathrm{X}}\{\hat{\phi}_n(x) - \phi(x)\} < -\eta \ i.o.) = 0 \ \forall \eta > 0$,

(ii) $\mathbf{P}(\sup_{x\in\mathbb{X}}\{\hat{\phi}_n(x)-\phi(x)\}>\eta \text{ i.o.})=0\ \forall\eta>0$,
(iii) $\sup_{x\in\mathbb{X}}\{|\hat{\phi}_n(x)-\phi(x)|\}\xrightarrow{a.s.}0$.

Step VII: We can now complete the proof of Theorem 3.1. Consider the class $\mathfrak{C}$ of all open rectangles $\mathcal{R}$ such that $\overline{\mathcal{R}}\subset\mathfrak{X}^\circ$ and whose vertices have rational coordinates. Then, $\mathfrak{C}$ is countable and $\bigcup_{\mathcal{R}\in\mathfrak{C}}\mathcal{R}=\mathfrak{X}^\circ$. Observe that Lemmas 3.2 and 3.3 imply that for any finite union $A:=\mathcal{R}_1\cup\cdots\cup\mathcal{R}_m$ of open rectangles $\mathcal{R}_1,\ldots,\mathcal{R}_m\in\mathfrak{C}$ there is, with probability 1, $n_0\in\mathbb{N}$ such that the sequence $(\hat{\phi}_n)_{n=n_0}^\infty$ is finite on $\mathrm{Conv}(A)$. From Lemma 3.9 we know that the least squares estimator converges at all rational points in $\mathfrak{X}^\circ$ with probability 1. Then, Theorem 10.8, page 90 of Rockafellar (1970) implies that (i) holds if $\mathfrak{X}^\circ$ is replaced by the convex hull of a finite union of rectangles belonging to $\mathfrak{C}$. Since there are countably many of such unions and any compact subset of $\mathfrak{X}^\circ$ is contained in one of these unions, we see that (i) holds. An application of Theorem 24.5, page 233 of Rockafellar (1970) on an open rectangle C containing x and satisfying $\overline{C}\subset\mathfrak{X}^\circ$ gives (ii) and (iii). Note that (iv) is a consequence of (iii). □

PROOF OF THEOREM 3.2. To prove the desired result we need the following lemma [whose proof is provided in Section 3 in Seijo and Sen (2010)] from convex analysis. The result is an extension of Theorem 25.7, page 248 of Rockafellar (1970), and might be of independent interest.

LEMMA 3.10. *Let $\mathcal{C}\subset\mathbb{R}^d$ be an open, convex set and f a convex function which is finite and differentiable on $\mathcal{C}$. Consider a sequence of convex functions $(f_n)_{n=1}^\infty$ which are finite on $\mathcal{C}$ and such that $f_n\to f$ pointwise on $\mathcal{C}$. Then, if $\mathbb{X}\subset\mathcal{C}$ is any compact set,*

$$\sup_{\substack{x\in\mathbb{X}\\ \xi\in\partial f_n(x)}}\{|\xi-\nabla f(x)|\}\to 0.$$

Defining the class $\mathfrak{C}$ of open rectangles as in the proof of Theorem 3.1, one can use a similar argument to obtain Theorem 3.2 from an application of Theorem 3.1 and the previous lemma. □

4. Proofs of some lemmas. Here we show some of the lemmas involved in the proof of the main theorem. We omit the proofs of Lemmas 3.2, 3.3, 3.4 and 3.10 since they are based on technical arguments from convex analysis and matrix algebra. They can be found in the supplement to this paper, Seijo and Sen (2010).

4.1. *Proof of Lemma* 3.1. We will first show that the event $[\inf_{x\in\mathbb{X}}\{\hat{\phi}_n(x)-\phi(x)\}\geq M \text{ i.o.}]$ has probability zero. Under this event, there is a subsequence

$(n_k)_{k=1}^{\infty}$ such that $\inf_{x\in\mathbb{X}}\{\hat{\phi}_{n_k}(x)-\phi(x)\}\geq M\ \forall k\in\mathbb{N}$. Then (10) implies that for this subsequence, with probability 1, we have

$$\overline{\lim_{k\to\infty}}\frac{1}{N_{n_k}(\mathbb{X})}\sum_{X_j\in\mathbb{X}}\{Y_j-\hat{\phi}_{n_k}(X_j)\}\leq -M. \tag{13}$$

On the other hand, it is seen [by solving the corresponding quadratic programming problems; see, e.g., Exercise 16.2, page 484 of Nocedal and Wright (1999)] that for any $\eta>0$, $m\in\mathbb{N}$,

$$\inf\left\{\frac{1}{m}\sum_{1\leq j\leq m}|\xi^j|^2:\frac{1}{m}\sum_{1\leq j\leq m}\xi^j\geq\eta,\xi\in\mathbb{R}^m\right\}=\eta^2, \tag{14}$$

$$\inf\left\{\frac{1}{m}\sum_{1\leq j\leq m}|\xi^j|^2:\frac{1}{m}\sum_{1\leq j\leq m}\xi^j\leq-\eta,\xi\in\mathbb{R}^m\right\}=\eta^2. \tag{15}$$

For $0<\delta<M$, using (15) with $\eta=M-\delta$ together with (12) and (13) we get that, with probability 1, we must have

$$\underline{\lim_{k\to\infty}}\frac{1}{n_k}\sum_{j=1}^{n_k}\bigl(Y_j-\hat{\phi}_{n_k}(X_j)\bigr)^2\geq\nu(\mathbb{X})(M-\delta)^2.$$

Letting $\delta\to0$, we actually get

$$\begin{aligned}&\underline{\lim_{k\to\infty}}\frac{1}{n_k}\sum_{j=1}^{n_k}\bigl(Y_j-\hat{\phi}_{n_k}(X_j)\bigr)^2\\&\quad\geq\nu(\mathbb{X})M^2>\sigma^2=\overline{\lim_{k\to\infty}}\frac{1}{n_k}\sum_{j=1}^{n_k}\bigl(Y_j-\phi(X_j)\bigr)^2\qquad\text{a.s.},\end{aligned}$$

which is impossible because $\hat{\phi}_{n_k}$ is the least squares estimator. Therefore,

$$\mathbf{P}\Bigl(\inf_{x\in\mathbb{X}}\{\hat{\phi}_n(x)-\phi(x)\}\geq M\text{ i.o.}\Bigr)=0.$$

A similar argument now using (14) gives

$$\mathbf{P}\Bigl(\sup_{x\in\mathbb{X}}\{\hat{\phi}_n(x)-\phi(x)\}\leq -M\text{ i.o.}\Bigr)=0,$$

which completes the proof of the lemma.

4.2. *Proof of Lemma 3.5.* The result is obvious for conditions (A1)–(A3) and (A5)–(A7) when $\sigma^2=0$. So we assume that $\sigma^2>0$ for (A1)–(A3) and (A5)–(A7). Let $\varepsilon>0$ and $M=\sup_{x\in\mathbb{X}}\{|x|\}$. Choose $\delta>0$ satisfying

$$\begin{aligned}&\frac{\varepsilon}{(2(2M+K\sqrt{d}+1)/n)\sum_{j=1}^n|Y_j-\phi(X_j)|}\\&\quad<\delta<\frac{\varepsilon}{((2M+K\sqrt{d}+1)/n)\sum_{j=1}^n|Y_j-\phi(X_j)|}\end{aligned} \tag{16}$$

for n large. Notice that δ is well defined and the quantity on the left is positive, finite and bounded away from 0 as $\underline{\lim}\,\frac{1}{n}\sum_{j=1}^{n}|Y_j-\phi(X_j)|>0$ a.s. under any set of regularity conditions [for (A2)–(A4), conditions (A4)(i) and (A4)(iii) imply that we can apply the version of the strong law of large numbers for uncorrelated random variables, as it appears in Chung (2001), page 108, Theorem 5.1.2, to the sequence $(|\varepsilon_j|)_{j=1}^{\infty}$; for (A1)–(A3) and (A5)–(A7) this is immediate as $\sigma^2>0$]. The definition of the class $\mathcal{D}_{K,\mathrm{X}}$ implies that all its members are Lipschitz functions with Lipschitz constant bounded by $K\sqrt{d}$, a consequence of Rockafellar (1970), Theorem 24.7, page 237. Hence, (16) implies that

$$\sup_{\substack{|x-y|<\delta\\ x,y\in\mathrm{X},\psi\in\mathcal{D}_{K,\mathrm{X}}}}\{|\psi(x)-\psi(y)|\}\le\frac{\varepsilon}{(1/n)\sum_{j=1}^{n}|Y_j-\phi(X_j)|}.$$

Now, define $N_n\in\mathbb{N}$ by $N_n=\lceil\frac{\operatorname{diam}(\mathrm{X})}{\delta}\rceil\vee\lceil\frac{2K\sqrt{d}}{\delta}\rceil$, where $\lceil\cdot\rceil$ denotes the ceiling function. Observe that (16) implies

$$N_n-1\le\big(\operatorname{diam}(\mathrm{X})\vee 2K\sqrt{d}\big)\frac{2(2M+K\sqrt{d}+1)}{\varepsilon}\left(\frac{1}{n}\sum_{j=1}^{n}|Y_j-\phi(X_j)|\right).\tag{17}$$

Then, we can divide the rectangles X and $[-K,K]^d$ in N_n^d subrectangles, all of which have diameters less than δ. In other words, we can write

$$[-K,K]^d=\bigcup_{1\le j\le N_n^d}R_j,$$

$$\mathrm{X}=\bigcup_{1\le j\le N_n^d}V_j$$

with $\operatorname{diam}(R_j)<\delta$ and $\operatorname{diam}(V_j)<\delta\ \forall j=1,\dots,N_n^d$. In the same way, we can divide the interval $[-K,K]$ in N_n subintervals $\mathcal{I}_1,\dots,\mathcal{I}_{N_n}$ each having length less than δ. For each $j=1,\dots,N_n^d$, let ξ_j and x_j be the centroids of R_j and V_j, respectively, and for $j=1,\dots,N_n$ let η_j be the midpoint of $\mathcal{I}_j$. Consider the class of functions $\mathcal{H}_{n,\varepsilon}$ defined by

$$\mathcal{H}_{n,\varepsilon}=\Big\{\max_{(s,t,j)\in\mathcal{S}}\{\langle\xi_s,\cdot-x_t\rangle+\eta_j\}:\mathcal{S}\subset\{1,\dots,N_n^d\}^2\times\{1,\dots,N_n\}\Big\}.$$

Observe that the number of elements in the class $\mathcal{H}_{n,\varepsilon}$ is bounded from above by $2^{N_n^{2d+1}}$. Now, take any $\psi\in\mathcal{D}_{K,\mathrm{X}}$. Pick any $\Xi_j\in\partial\psi(X_j)$. Then, for any j such that $X_j\in\mathrm{X}$, there are $s_j,t_j\in\{1,\dots,N_n^d\}$ and $\tau_j\in\{1,\dots,N_n\}$ such that $|\Xi_j-\xi_{s_j}|$, $|X_j-x_{t_j}|$ and $|\psi(x_{t_j})-\eta_{\tau_j}|$ are all less than δ. We then have that

$$\begin{aligned}&\sup_{x\in\mathrm{X}}\big\{\big|\langle\xi_{s_j},x-x_{t_j}\rangle+\eta_{\tau_j}-\big(\langle\Xi_j,x-X_j\rangle+\psi(X_j)\big)\big|\big\}\\&\qquad\le 2M|\xi_{s_j}-\Xi_j|+K\sqrt{d}|x_{t_j}-X_j|+\delta<\big(2M+K\sqrt{d}+1\big)\delta\end{aligned}\tag{18}$$

by an application of the Cauchy–Schwarz inequality. But then, (16) implies that if we define the functions $\tilde{\psi}$ and g as

$$\tilde{\psi}(x) = \max_{X_j \in \mathbb{X}} \{\langle \Xi_j, x - X_j \rangle + \psi(X_j)\},$$

$$g(x) = \max_{X_j \in \mathbb{X}} \{\langle \xi_{s_j}, x - x_{t_j} \rangle + \eta_{\tau_j}\},$$

then we have

$$\tilde{\psi}(X_j) = \psi(X_j) \qquad \text{for } j \text{ such that } X_j \in \mathbb{X}, \tag{19}$$

$$\|g - \tilde{\psi}\|_{\mathbb{X}} \le \frac{\varepsilon}{(1/n)\sum_{j=1}^{n} |Y_j - \phi(X_j)|} \qquad [\text{from (18)}], \tag{20}$$

$$g \in \mathcal{H}_{n,\varepsilon}. \tag{21}$$

Note that (19) follows from the definition of subgradients. All these facts put together give that for any $f(x, y) = \psi(x)(y - \phi(x)) \in \mathcal{G}_{K,\mathbb{X}}$, $\psi \in \mathcal{D}_{K,\mathbb{X}}$, there is $g \in \mathcal{H}_{n,\varepsilon}$ such that

$$\int_{\mathbb{X}} |f(x, y) - g(x)(y - \phi(x))| \mu_n(dx, dy) < \varepsilon$$

and hence

$$N(\varepsilon, \mathcal{G}_{K,\mathbb{X}}, \mathbb{L}_1(\mathbb{X} \times \mathbb{R}, \mu_n)) \le \#\mathcal{H}_{n,\varepsilon} \le 2^{N_n^{2d+1}}.$$

But then, the strong law of large numbers and (17) give that $\overline{\lim} N_n < \infty$ a.s. Furthermore, by replacing ε with $\frac{\varepsilon}{n}\sum_{j=1}^{n} |Y_j - \phi(X_j)|$ in the entire construction just made, we can see that the covering numbers $N(\frac{\varepsilon}{n}\sum_{j=1}^{n} |Y_j - \phi(X_j)|, \mathcal{G}_{K,\mathbb{X}}, \mathbb{L}_1(\mathbb{X} \times \mathbb{R}, \mu_n))$ depend neither on the Y's nor on ϕ. Taking $B_\varepsilon = (\operatorname{diam}(\mathbb{X}) \vee K\sqrt{d}) \frac{2(2M + K\sqrt{d}+1)}{\varepsilon} + 1$ and $A_\varepsilon = 2^{B_\varepsilon^{2d+1}}$, it is seen that the second part of the result holds.

4.3. *Proof of Lemma 3.6.* Note that for every m, we have

$$\frac{1}{n_k} \sum_{1 \le j \le n_k} \mathbf{E}(\varepsilon_j^2) \le \frac{1}{n_k} \sum_{\substack{X_j \in \mathbb{X}_m \\ 1 \le j \le n_k}} \mathbf{E}(\varepsilon_j^2) + \frac{N_{n_k}(\mathfrak{X} \setminus \mathbb{X}_m)}{n_k} \sup_{j \in \mathbb{N}} \{\mathbf{E}(\varepsilon_j^2)\}.$$

Taking limit inferior on both sides as $k \to \infty$, we get

$$\sigma^2 \le \varliminf_{k \to \infty} \frac{1}{n_k} \sum_{\substack{X_j \in \mathbb{X}_m \\ 1 \le j \le n_k}} \mathbf{E}(\varepsilon_j^2) + \nu(\mathfrak{X} \setminus \mathbb{X}_m) \sup_{j \in \mathbb{N}} \{\mathbf{E}(\varepsilon_j^2)\}.$$

Now taking the limit as $m \to \infty$ we get the result because the opposite inequality is trivial.

4.4. *Proof of Lemma 3.7.* We may assume that $\mathbb{X}$ is a compact rectangle. Here we need to make a distinction between the design schemes. In the case of the

stochastic design, the proof is an immediate consequence of Lemma 3.5 and Theorem 2.4.3, page 123 of Van der Vaart and Wellner (1996). Thus, we focus on the fixed design scenario.

For notational convenience, we write $M = \sup_{j\in\mathbb{N}}\{\mathbf{E}(\varepsilon_j^2)\}$ and $\sum_{X_j\in\mathbb{X}}$ instead of the more cumbersome $\sum_{1\le j\le n\,:\,X_j\in\mathbb{X}}$. Letting $\varepsilon_j = Y_j - \phi(X_j)$ (and using the same notation as in the proof of Lemma 3.7), first observe that the random quantity

$$\sup_{\psi\in\mathcal{D}_{K,\mathbb{X}}}\left\{\left|\frac{1}{n}\sum_{\{X_j\in\mathbb{X}\}}\psi(X_j)\varepsilon_j\right|\right\} = \sup_{m\in\mathbb{N}}\left\{\sup_{g\in\mathcal{H}_{n,1/m}}\left\{\left|\frac{1}{n}\sum_{\{X_j\in\mathbb{X}\}}g(X_j)\varepsilon_j\right|\right\}\right\}$$

by (19), (20) and (21) and is thus measurable.

All of the following arguments are valid for both (A1)–(A3) and (A2)–(A4). Lyapunov's inequality (which states that for any random variable X and $1 \le p \le q \le \infty$ we have $\|X\|_p \le \|X\|_q$) and the strong law of large numbers imply

$$\overline{\lim_{m\to\infty}}\frac{1}{m}\sum_{1\le j\le m}|\varepsilon_j| = \overline{\lim_{m\to\infty}}\frac{1}{m}\sum_{1\le j\le m}\mathbf{E}(|\varepsilon_j|) \le \sqrt{M} \qquad \text{a.s.} \tag{22}$$

Let $\eta > 0$. From Lemma 3.5 we know that the covering numbers $a_n := N(\frac{\eta}{n}\sum_{j=1}^n|Y_j - \phi(X_j)|, \mathcal{G}_{K,\mathbb{X}}, \mathbb{L}_1(\mathbb{X}\times\mathbb{R},\mu_n))$ are not random and uniformly bounded by a constant A_η. Therefore, for any $n\in\mathbb{N}$ we can find a class $\mathcal{A}_n \subset \mathcal{D}_{K,\mathbb{X}}$ with exactly a_n elements such that $\{\psi(x)(y-\phi(x))\}_{\psi\in\mathcal{A}_n}$ forms an $(\frac{\eta}{n}\sum_{j=1}^n|Y_j - \phi(X_j)|)$-net for $\mathcal{G}_{K,\mathbb{X}}$ with respect to $\mathbb{L}_1(\mathbb{X}\times\mathbb{R},\mu_n)$. It follows that

$$\sup_{\psi\in\mathcal{D}_{K,\mathbb{X}}}\left\{\left|\frac{1}{n}\sum_{X_j\in\mathbb{X}}\psi(X_j)\varepsilon_j\right|\right\} \le \frac{\eta}{n}\sum_{1\le j\le n}|\varepsilon_j| + \sup_{\psi\in\mathcal{A}_n}\left\{\left|\frac{1}{n}\sum_{X_j\in\mathbb{X}}\psi(X_j)\varepsilon_j\right|\right\}. \tag{23}$$

With (23) in mind, we make the following definitions:

$$B_n = \sup_{\psi\in\mathcal{A}_n}\left\{\left|\frac{1}{n}\sum_{X_j\in\mathbb{X}}\psi(X_j)\varepsilon_j\right|\right\},$$

$$C_n = \sup_{\psi\in\mathcal{A}_n}\left\{\left|\frac{1}{n}\sum_{1\le j\le\lfloor\sqrt{n}\rfloor^2\,:\,X_j\in\mathbb{X}}\psi(X_j)\varepsilon_j\right|\right\},$$

$$D_n = \sup_{\substack{\psi\in\mathcal{A}_k\\ n^2\le k<(n+1)^2}}\left\{\left|\frac{1}{k}\sum_{n^2<j\le k\,:\,X_j\in\mathbb{X}}\psi(X_j)\varepsilon_j\right|\right\},$$

where $\lfloor\cdot\rfloor$ denotes the floor function. Now, pick $\delta > 0$ and observe that

$$\mathbf{P}(B_n > \delta) = \mathbf{P}\Bigg(\bigcup_{\psi\in\mathcal{A}_n}\left[\left|\sum_{X_j\in\mathbb{X}}\psi(X_j)\varepsilon_j\right| > n\delta\right]\Bigg)$$

$$\le \sum_{\psi\in\mathcal{A}_n}\frac{1}{n^2\delta^2}M\sum_{X_j\in\mathbb{X}}\psi(X_j)^2 \le \frac{K^2MA_\eta}{n\delta^2}.$$

The Borel–Cantelli lemma then implies that $\mathbf{P}(B_{n^2} > \delta \text{ i.o.}) = 0$. Letting $\delta \to 0$ through a decreasing sequence gives

$$B_{n^2} \xrightarrow{\text{a.s.}} 0. \tag{24}$$

On the other hand, the definition of C_n implies that

$$C_n \le \frac{\lfloor \sqrt{n} \rfloor^2}{n} B_{\lfloor \sqrt{n} \rfloor^2} + \frac{\eta}{n} \sum_{1 \le j \le \lfloor \sqrt{n} \rfloor^2} |\varepsilon_j|, \tag{25}$$

which together with (24) and (22) gives

$$\overline{\lim}\, C_n \le \eta \sqrt{M} \qquad \text{almost surely.} \tag{26}$$

Note that (25) is a consequence of the fact that for any $\psi \in \mathcal{A}_n$, there exists $g \in \mathcal{A}_{\lfloor \sqrt{n} \rfloor^2}$ such that if $\mathcal{J}_n = \{1 \le j \le \lfloor \sqrt{n} \rfloor^2 : X_j \in \mathbb{X}\}$, then

$$\begin{aligned} \left| \frac{1}{n} \sum_{j \in \mathcal{J}_n} \psi(X_j) \varepsilon_j \right| &\le \left| \frac{1}{n} \sum_{j \in \mathcal{J}_n} \big(\psi(X_j) - g(X_j)\big) \varepsilon_j \right| + \left| \frac{1}{n} \sum_{j \in \mathcal{J}_n} g(X_j) \varepsilon_j \right| \\ &\le \left(\frac{\lfloor \sqrt{n} \rfloor^2}{n} \right) \frac{\eta}{\lfloor \sqrt{n} \rfloor^2} \sum_{1 \le j \le \lfloor \sqrt{n} \rfloor^2} |\varepsilon_j| + \frac{\lfloor \sqrt{n} \rfloor^2}{n} B_{\lfloor \sqrt{n} \rfloor^2}. \end{aligned}$$

Now, an argument similar to the one used in (24) gives

$$\begin{aligned} \mathbf{P}(D_n > \delta) &= \mathbf{P}\Bigg(\bigcup_{\substack{\psi \in \mathcal{A}_k \\ n^2 \le k < (n+1)^2}} \left[\left| \sum_{n^2 < j \le k \,:\, X_j \in \mathbb{X}} \psi(X_j) \varepsilon_j \right| > k\delta \right] \Bigg) \\ &\le \sum_{\substack{\psi \in \mathcal{A}_k \\ n^2 \le k < (n+1)^2}} \mathbf{P}\left(\left| \sum_{n^2 < j \le k \,:\, X_j \in \mathbb{X}} \psi(X_j) \varepsilon_j \right| > k\delta \right) \qquad (27) \\ &\le \sum_{\substack{\psi \in \mathcal{A}_k \\ n^2 \le k < (n+1)^2}} \frac{K^2 M (k - n^2)}{k^2 \delta^2} \le \frac{K^2 M A_\eta (2n+1)^2}{n^4 \delta^2}. \end{aligned}$$

Again, one can use (27) and the Borel–Cantelli lemma to prove that $\mathbf{P}(D_n > \delta \text{ i.o.}) = 0$ and then let $\delta \to 0$ through a decreasing sequence to obtain

$$D_n \xrightarrow{\text{a.s.}} 0. \tag{28}$$

Finally, one sees that

$$\sup_{\psi \in \mathcal{A}_n} \left\{ \left| \frac{1}{n} \sum_{X_j \in \mathbb{X}} \psi(X_j) \big(Y_j - \phi(X_j)\big) \right| \right\} = B_n \le C_n + D_{\lfloor \sqrt{n} \rfloor},$$

which combined with (26) and (28) gives

$$\overline{\lim}\, B_n \le \eta\sqrt{M} \qquad \text{almost surely.}$$

Taking (23) into account we get

$$\overline{\lim_{n\to\infty}} \sup_{\psi\in\mathcal{D}_{K,\mathtt{X}}} \left\{ \left| \frac{1}{n} \sum_{1\le j\le n:\, X_j\in\mathtt{X}} \psi(X_j)\big(Y_j - \phi(X_j)\big) \right| \right\} \le 2\eta\sqrt{M} \qquad \text{almost surely.}$$

Letting $\eta \to 0$, we get the desired result.

4.5. *Proof of Lemma 3.8.* We can assume, without loss of generality, that $\mathtt{X}$ is a finite union of compact rectangles. Consider a sequence $(\mathtt{X}_m)_{m=1}^{\infty}$ satisfying the following properties:

(a) $\mathtt{X} \subset \mathtt{X}_m \subset \mathfrak{X}^\circ$ $\forall m \in \mathbb{N}$.
(b) $\nu(\mathtt{X}_m) > 1 - \frac{1}{m}$ $\forall m \in \mathbb{N}$.
(c) $\mathtt{X}_m \subset \mathtt{X}_{m+1}$ $\forall m \in \mathbb{N}$.
(d) Every $\mathtt{X}_m$ can be expressed as a finite union of compact rectangles with positive Lebesgue measure.

The existence of such a sequence follows from the inner regularity of Borel probability measures on $\mathbb{R}^d$ and from the fact that since $\mathfrak{X}^\circ$ is open, for any compact set $F \subset \mathfrak{X}^\circ$ we can find a finite cover composed by compact rectangles with positive Lebesgue measure and completely contained in $\mathfrak{X}^\circ$. Also, from Lemmas 3.2, 3.3 and 3.4 and the fact that $\mathfrak{X} \subset \mathrm{Dom}(\phi)$, for any $m \in \mathbb{N}$ we can find $K_m > 0$ such that

$$\|\phi\|_{\mathtt{X}_m} \le K_m \quad \text{and} \quad \mathbf{P}(\|\hat{\phi}_n\|_{\mathtt{X}_m} > K_m \text{ i.o.}) = 0; \tag{29}$$

$$\sup_{\substack{x\in\mathtt{X}_m \\ \xi\in\partial\phi(x)}} \{|\xi|\} \le K_m \quad \text{and} \quad \mathbf{P}^*\Big(\sup_{\substack{x\in\mathtt{X}_m \\ \xi\in\partial\hat{\phi}_n(x)}} \{|\xi|\} > K_m \text{ i.o.}\Big) = 0. \tag{30}$$

Fix $\eta > 0$ and consider the sets

$$A = \Big[\inf_{x\in\mathtt{X}}\{\phi(x) - \hat{\phi}_n(x)\} \ge \eta \text{ i.o.}\Big],$$

$$B = [\|\hat{\phi}_n\|_{\mathtt{X}_m} \le K_m \text{ a.a.}],$$

$$C = \Big[\sup_{\substack{x\in\mathtt{X}_m \\ \xi\in\partial\hat{\phi}_n(x)}} \{|\xi|\} \le K_m \text{ a.a.}\Big].$$

Suppose now that $A \cap B \cap C$ is known to be true. Then, there is a subsequence $(n_k)_{k=1}^{\infty}$ such that $\inf_{x\in\mathtt{X}}\{\phi(x) - \hat{\phi}_{n_k}(x)\} \ge \eta$ $\forall k \in \mathbb{N}$ and $\frac{1}{n_k}\sum_{j=1}^{n_k} \mathbf{E}(\varepsilon_j^2) \to \sigma^2$.

Taking (29) and (30) into account, we have that for k large enough the inequality

$$\begin{aligned}&\frac{1}{n_k}\sum_{j=1}^{n_k}\bigl(Y_j-\hat{\phi}_{n_k}(X_j)\bigr)^2\\&\quad\geq\frac{1}{n_k}\sum_{X_j\in\mathbb{X}_m}\bigl(Y_j-\phi(X_j)\bigr)^2+\frac{2}{n_k}\sum_{X_j\in\mathbb{X}_m}\bigl(Y_j-\phi(X_j)\bigr)\bigl(\phi(X_j)-\hat{\phi}_{n_k}(X_j)\bigr)\\&\qquad+\frac{1}{n_k}\sum_{X_j\in\mathbb{X}_m}\bigl(\phi(X_j)-\hat{\phi}_{n_k}(X_j)\bigr)^2\end{aligned}$$

implies

$$\begin{aligned}&\frac{1}{n_k}\sum_{j=1}^{n_k}\bigl(Y_j-\hat{\phi}_{n_k}(X_j)\bigr)^2\\&\quad\geq\frac{1}{n_k}\sum_{X_j\in\mathbb{X}_m}\bigl(Y_j-\phi(X_j)\bigr)^2+\frac{N_{n_k}(\mathbb{X})}{n_k}\eta^2\\&\qquad-4\sup_{\psi\in\mathcal{D}_{K_m,\mathbb{X}_m}}\biggl\{\biggl|\frac{1}{n_k}\sum_{\{1\leq j\leq n_k\,:\,X_j\in\mathbb{X}_m\}}\psi(X_j)\bigl(Y_j-\phi(X_j)\bigr)\biggr|\biggr\}.\end{aligned}$$

Thus, from Lemma 3.7 we can conclude that

$$\varliminf_{k\to\infty}\frac{1}{n_k}\sum_{1\leq j\leq n_k}\bigl(Y_j-\hat{\phi}_{n_k}(X_j)\bigr)^2\geq\nu(\mathbb{X}_m)\sigma^2+\nu(\mathbb{X})\eta^2\qquad\text{if (A1)–(A3) hold.}$$

Under (A2)–(A4) and (A5)–(A7) the left-hand side of the last display is bounded from below by

$$\varliminf_{k\to\infty}\frac{1}{n_k}\sum_{X_j\in\mathbb{X}_m}\bigl(Y_j-\phi(X_j)\bigr)^2+\nu(\mathbb{X})\eta^2$$

and

$$\int_{\mathbb{X}_m}\bigl(y-\phi(x)\bigr)^2\mu(dx,dy)+\nu(\mathbb{X})\eta^2,$$

respectively.

Finally, using (a)–(d), the strong law of large numbers [for (A2)–(A4) we can apply a version of the strong law of large numbers for independent random variables thanks to condition (A4)(ii); see Williams (1991), Lemma 12.8, page 118, or Folland (1999), Theorem 10.12, page 322] and Lemma 3.6 we can let $m\to\infty$ to see that, under any of (A1)–(A3), (A2)–(A4) or (A5)–(A7),

$$\varliminf_{k\to\infty}\frac{1}{n_k}\sum_{1\leq j\leq n_k}\bigl(Y_j-\hat{\phi}_{n_k}(X_j)\bigr)^2\geq\sigma^2+\nu(\mathbb{X})\eta^2,$$

which is impossible because $\hat{\phi}_{n_k}$ is the least squares estimator.

Therefore $\mathbf{P}^*(A \cap B \cap C) = 0$ and, since $\mathbf{P}_*(B \cap C) = 1$,

$$\mathbf{P}(A) = \mathbf{P}\Big(\inf_{x \in \mathbb{X}} \{\phi(x) - \hat{\phi}_n(x)\} \geq \eta \text{ i.o.}\Big) = 0.$$

This finishes the proof of (i). The second assertion follows from similar arguments.

4.6. *Proof of Lemma 3.9.* We can assume, without loss of generality, that $\mathbb{X}$ is a finite union of compact rectangles. Pick $K_{\mathbb{X}}$ such that

$$\sup_{\substack{x \in \mathbb{X} \\ \xi \in \partial\phi(x)}} \{|\xi|\} \leq K_{\mathbb{X}} \quad \text{and} \quad \mathbf{P}^*\Big(\sup_{\substack{x \in \mathbb{X} \\ \xi \in \partial\hat{\phi}_n(x)}} \{|\xi|\} > K_{\mathbb{X}} \text{ i.o.}\Big) = 0.$$

Let $\eta > 0$ and $\delta = \frac{\eta}{3K_{\mathbb{X}}}$. We can then divide $\mathbb{X}$ in M subrectangles $\{\mathcal{C}_1, \ldots, \mathcal{C}_M\}$ all having diameter less than δ. Define the events

$$A = \Big[\bigcap_{1 \leq k \leq M} \inf_{x \in \mathcal{C}_k} \{\hat{\phi}_n(x) - \phi(x)\} < \frac{\eta}{3} \text{ a.a.} \Big],$$

$$B = \Big[\sup_{\substack{x \in \mathbb{X} \\ \xi \in \partial\hat{\phi}_n(x)}} \{|\xi|\} \leq K_{\mathbb{X}} \text{ a.a.} \Big].$$

We will show that $A \cap B \subset [\sup_{x \in \mathbb{X}} \{\hat{\phi}_n(x) - \phi(x)\} \leq \eta \text{ a.a.}]$. Suppose $A \cap B$ is true. Then, there is $N \in \mathbb{N}$ such that for any $n \geq N$ we can find $\Xi_{n,k} \in \mathcal{C}_k$ such that $\hat{\phi}_n(\Xi_{n,k}) - \phi(\Xi_{n,k}) < \frac{\eta}{3}$. Moreover, we can make N large enough such that for any $n \geq N$, $K_{\mathbb{X}}$ is an upper bound for all the subgradients of $\hat{\phi}_n$ on $\mathbb{X}$. Then, for any $\xi \in \mathcal{C}_k$ we obtain from the Lipschitz property

$$\begin{aligned} \hat{\phi}_n(\xi) - \phi(\xi) &= \big(\hat{\phi}_n(\Xi_{n,k}) - \phi(\Xi_{n,k})\big) + \big(\phi(\Xi_{n,k}) - \phi(\xi)\big) \\ &\quad + \big(\hat{\phi}_n(\xi) - \hat{\phi}_n(\Xi_{n,k})\big) \\ &\leq \frac{\eta}{3} + K_{\mathbb{X}}\delta + K_{\mathbb{X}}\delta \leq \eta. \end{aligned}$$

Therefore,

$$\sup_{x \in \mathcal{C}_k} \{\hat{\phi}_n(x) - \phi(x)\} \leq \eta \qquad \forall 1 \leq k \leq M, \forall n \geq N,$$

which implies

$$\sup_{x \in \mathbb{X}} \{\hat{\phi}_n(x) - \phi(x)\} \leq \eta \qquad \forall n \geq N.$$

Considering Lemmas 3.8(ii) and 3.4; $A \cap B \subset [\sup_{x \in \mathbb{X}} \{\hat{\phi}_n(x) - \phi(x)\} \leq \eta \text{ a.a.}]$ and $\mathbf{P}_*(A \cap B) = 1$ we obtain (ii). The first assertion follows from similar arguments and (iii) is a direct consequence of (i) and (ii).

Acknowledgments. The authors are very grateful to the reviewers for their constructive and useful comments which helped to improve the presentation of the paper.

SUPPLEMENTARY MATERIAL

Supplement to "Nonparametric least squares estimation of a multivariate convex regression function" (DOI: 10.1214/10-AOS852SUPP; .pdf). The supplementary file contains the proofs of some technical results that were omitted from the main draft due to their length.

REFERENCES

ALLON, G., BEENSTOCK, M., HACKMAN, S., PASSY, U. and SHAPIRO, A. (2007). Nonparametric estimation of concave production technologies by entropic methods. *J. Appl. Econometrics* **22** 795–816. MR2370975

BANKER, R. D. and MAINDIRATTA, A. (1992). Maximum likelihood estimation of monotone and concave production frontiers. *J. Productiv. Anal.* **3** 401–415.

BERESTEANU, A. (2007). Nonparametric estimation of regression functions under restrictions on partial derivatives. Available at http://www.pitt.edu/~arie/shape.pdf.

BIRKE, M. and DETTE, H. (2007). Estimating a convex function in nonparametric regression. *Scand. J. Stat.* **34** 384–404. MR2346646

BRONŠTEĬN, E. M. (1978). Extremal convex functions. *Sibirsk. Mat. Zh.* **19** 10–18. MR0482540

BRUNK, H. D. (1955). Maximum likelihood estimates of monotone parameters. *Ann. Math. Statist.* **26** 607–616. MR0073894

BRUNK, H. D. (1970). Estimation of isotonic regression. In *Nonparametric Techniques in Statistical Inference* 177–197. Cambridge Univ. Press, New York. MR0277070

CHUNG, K. L. (2001). *A Course in Probability Theory*. Academic Press, San Diego, CA. MR1796326

CONWAY, J. (1985). *A Course in Functional Analysis*. Springer, New York. MR0768926

CULE, M. and SAMWORTH, R. (2010). Theoretical properties of the log-concave maximum likelihood estimator of a multidimensional density. *Electron. J. Stat.* **4** 254–270. MR2645484

CULE, M., SAMWORTH, R. and STEWART, M. (2010). Maximum likelihood estimation of a multidimensional log-concave density. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **72** 545–607.

DUDLEY, R. M. (1977). On second derivatives of convex functions. *Math. Scand.* **41** 159–174. MR0482164

FOLLAND, G. (1999). *Real Analysis*: *Modern Techniques and Their Applications*. Wiley, New York. MR1681462

GRENANDER, U. (1956). On the theory of mortality measurement. II. *Skand. Aktuarietidskr.* **39** 125–153. MR0093415

GROENEBOOM, P., JONGBLOED, G. and WELLNER, J. (2001). Estimation of a convex function: Characterizations and asymptotic theory. *Ann. Statist.* **29** 1653–1698. MR1891742

HANSON, D. L. and PLEDGER, G. (1976). Consistency in concave regression. *Ann. Statist.* **4** 1038–1050. MR0426273

HILDRETH, C. (1954). Point estimates of ordinates of concave functions. *J. Amer. Statist. Assoc.* **49** 598–619. MR0065093

JOHANSEN, S. (1974). The extremal convex functions. *Math. Scand.* **41** 61–68. MR0346517

KUOSMANEN, T. (2008). Representation theorem for convex nonparametric least squares. *Econom. J.* **11** 308–325.

LUENBERGER, D. (1984). *Linear and Nonlinear Programming*. Addison-Wesley, Reading, MA.
MAMMEN, E. (1991). Nonparametric regression under qualitative smoothness assumptions. *Ann. Statist.* **19** 741–759. MR1105842
MATZKIN, R. L. (1991). Semiparametric estimation of monotone concave utility functions for polychotomous choice models. *Econometrica* **59** 1351–1327. MR1133036
MATZKIN, R. L. (1993). Nonparametric identification and estimation of polychotomous choice models. *J. Econometrics* **58** 137–168. MR1230983
NOCEDAL, J. and WRIGHT, S. (1999). *Numerical Optimization*. Springer, New York. MR1713114
ROCKAFELLAR, T. R. (1970). *Convex Analysis*. Princeton Univ. Press, Princeton, NJ. MR0274683
SARATH, B. and MAINDIRATTA, A. (1997). On the consistency of maximum likelihood estimation of monotone and concave production frontiers. *J. Productiv. Anal.* **8** 239–246.
SCHUHMACHER, D. and DÜMBGEN, L. (2010). Consistency of multivariate log-concave density estimators. *Statist. Probab. Lett.* **80** 376–380. MR2593576
SCHUHMACHER, D., HÜSLER, A. and DÜMBGEN, L. (2009). Multivariate log-concave distributions as a nearly parametric model. Technical report, Univ. Bern. Available at http://arxiv.org/abs/0907.0250.
SEIJO, E. and SEN, B. (2011). Supplement to "Nonparametric least squares estimation of a multivariate convex regression function." DOI:10.1214/10-AOS852SUPP.
SEREGIN, A. and WELLNER, J. (2010). Nonparametric estimation of multivariate convex-transformed densities. *Ann. Statist.* **38** 3751–3781. MR2766867
VAN DER VAART, A. and WELLNER, J. (1996). *Weak Convergence and Empirical Processes*. Springer, New York. MR1385671
VARIAN, H. (1982). The nonparametric approach to demand analysis. *Econometrica* **50** 945–973. MR0666119
VARIAN, H. (1984). The nonparametric approach to production analysis. *Econometrica* **52** 579–597. MR0740302
WILLIAMS, D. (1991). *Probability with Martingales*. Cambridge Univ. Press, Cambridge. MR1155402
ZHANG, C. H. (2002). Risk bounds in isotonic regression. *Ann. Statist.* **30** 528–555. MR1902898

DEPARTMENT OF STATISTICS
COLUMBIA UNIVERSITY
1032 AMSTERDAM AVENUE
NEW YORK, NEW YORK 10027
USA
E-MAIL: emilio@stat.columbia.edu
bodhi@stat.columbia.edu
URL: http://www.stat.columbia.edu/~emilio
http://www.stat.columbia.edu/~bodhi