International Journal of Statistics and Probability; Vol. 10, No. 1; January 2021
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Consistency of Penalized Convex Regression

Eunji Lim

Correspondence: Eunji Lim, Department of Decision Sciences and Marketing, School of Business, Adelphi University, Garden City, NY 11530, USA. E-mail: elim@adelphi.edu

Received: November 1, 2020 Accepted: December 4, 2020 Online Published: December 20, 2020

doi:10.5539/ijsp.v10n1p69 URL: https://doi.org/10.5539/ijsp.v10n1p69

Abstract

We consider the problem of estimating an unknown convex function $f_* : (0,1)^d \to \mathbb{R}$ from data $(X_1, Y_1), \cdots, (X_n, Y_n)$. A simple approach is finding a convex function that is the closest to the data points by minimizing the sum of squared errors over all convex functions. The convex regression estimator, which is computed this way, suffers from a drawback of having extremely large subgradients near the boundary of its domain. To remedy this situation, the penalized convex regression estimator, which minimizes the sum of squared errors plus the sum of squared norms of the subgradient over all convex functions, is recently proposed. In this paper, we prove that the penalized convex regression estimator and its subgradient converge with probability one to f_* and its subgradient, respectively, as $n \to \infty$, and hence, establish the legitimacy of the penalized convex regression estimator.

Keywords: convexity regression, penalized convex regression

1. Introduction

We consider the problem of estimating an unknown function $f_* : (0,1)^d \to \mathbb{R}$ from noisy observations $(X_1, Y_1), \cdots (X_n, Y_n)$ when one cannot assume any parametric form on f_* and the only available information is the fact that f_* is convex. We assume that the X_i's are independent and identically distributed (iid) $(0,1)^d$-valued random vectors, and

$$Y_i = f_*(X_i) + \varepsilon_i$$

for $1 \le i \le n$, where the ε_i's are iid random variables with mean zero and a finite variance.

This situation arises in many practical settings. For example, the long run average waiting time per customer in a single server queue is proven to be convex in the service rate (Weber (1983)). Various examples exist in the context of economics and queueing systems.

When the only available information is the fact that f_* is convex, a simple approach to estimating f_* is finding a convex function that is the closest to the data set $(X_1, Y_1), \cdots (X_n, Y_n)$. In other words, we seek to find the solution to the following problem:

$$\begin{aligned} \text{Minimize} \quad & \frac{1}{n}\sum_{i=1}^{n}(Y_i - f(X_i))^2 \\ \text{Subject to} \quad & f : (0,1)^d \to \mathbb{R} \text{ is convex.} \end{aligned} \tag{1}$$

It appears that (1) is an infinite-dimensional problem. However, we can notice that there is a convex function $f : (0,1)^d \to \mathbb{R}$ passing through $(X_1, f_1), \cdots, (X_n, f_n)$ if and only if there exists a subgradient $\xi_i \in \mathbb{R}^d$ at each X_i for $1 \le i \le n$, satisfying

$$f_j \ge f_i + \xi_i^T (X_j - X_i)$$

for $1 \le j \le n$; see pp. 337–338 of Boyd and Vandenberghe (2004). Using this fact, it can be seen that (1) is equivalent to the following finite–dimensional problem:

$$\begin{aligned} \text{Minimize} \quad & \frac{1}{n}\sum_{i=1}^{n}(Y_i - f_i)^2 \\ \text{Subject to} \quad & f_j \ge f_i + \xi_i^T (X_j - X_i), \quad 1 \le i, j \le n, \end{aligned} \tag{2}$$

where $f_1, \cdots, f_n \in \mathbb{R}$ and $\xi_1, \cdots, \xi_n \in \mathbb{R}^d$; see Hildreth (1954), Kuosmanen (2009), and Seijo and Sen (2011) for the details. In (2), f_i corresponds to the value of the fitted function at X_i, and ξ_i is a subgradient of the fitted function at X_i for $1 \le i \le n$.

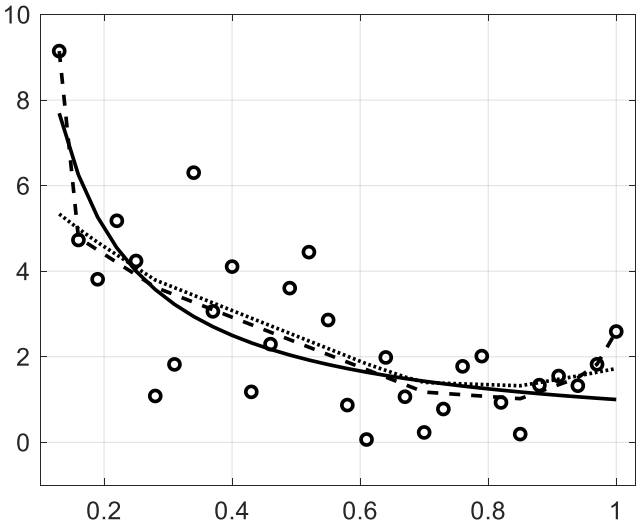

Figure 1. $X_i = 0.1 + 3i/100$ for $1 \le i \le 30$. The solid line is the unknown function f_* defined by $f_*(x) = 1/x$ for $x \in (0, 1)$ and the circles are the observations $Y_i = f_*(X_i) + \varepsilon_i$, the dashed line is the convex regression estimator, and the dotted line is the penalized convex regression estimator with $\lambda_n = 0.01$. The ε_i's follow iid standard normal distributions

We refer to the solution to (2) as the convex regression estimator, and this estimator has gained a great deal of attention from numerous researchers. Hanson and Pledger (1976) established consistency for the case when $d = 1$, Groeneboom et al. (2001) computed the rate of convergence for the case when $d = 2$, Seijo and Sen (2011) studies consistency for the case when $d > 1$, and Mazumder et al. (2019) proposed an efficient algorithm for solving (2). One can note that (2) is a convex quadratic program with $n(d + 1)$ decision variables and n^2 linear constraints, so one can solve (2) by using convex programming solvers.

One of the drawbacks of the convex regression estimator is that it tends to overfit the data set near the boundary of the domain, so its subgradient gets large near the boundary. The main reason of this undesirable situation is that (2) is formulated in a way that only the sum of squared errors is minimized. Thus, one way to remedy this situation is adding a penalty term to the objective function of (2), which leads to the following formulation:

$$\begin{aligned} \text{Minimize} \quad & \frac{1}{n}\sum_{i=1}^{n}(Y_i - f_i)^2 + \frac{\lambda_n}{n}\sum_{i=1}^{n}\|\xi_i\|^2 \\ \text{Subject to} \quad & f_j \ge f_i + \xi_i^T(X_j - X_i), \quad 1 \le i, j \le n, \end{aligned} \tag{3}$$

where $f_1, \cdots, f_n \in \mathbb{R}$, $\xi_1, \cdots, \xi_n \in \mathbb{R}^d$, $\lambda_n \ge 0$ is the smoothing constant, and $\|(z_1, \cdots, z_d)\| \triangleq (z_1^2 + \cdots + z_d^2)^{1/2}$ for $(z_1, \cdots, z_d) \in \mathbb{R}^d$; see Chen et al. (2020) for the formulation and Bertsimas and Mundru (2020) for an efficient numerical algorithm that solves (3).

The solution to (3), which we refer to as the "penalized convex regression estimator," exhibits nice numerical behavior such as bounded subgradients near the boundary of the domain. For example, Figure 1 shows an instance of the penalized convex regression estimator, compared to that of the convex regression estimator. The figure shows how the convex regression estimator overfits data near the boundary of the domain, forcing the subgradient to be large, whereas the penalized convex regression estimator has bounded subgradients throughout the domain. Thus, when estimating both f_* and its subgradient, one may prefer the penalized convex regression estimator. Despite its appealing numerical performance, the statistical foundation of the penalized convex regression estimator has not been established so far.

The goal of this paper is to establish strong consistency of the penalized least squares estimator and its subgradient uniformly over any compact subset of $(0, 1)^d$. Specifically, Theorems 1 and 2 state that the penalized least squares estimator and its subgradient converge almost surely to f_* and the subgradient of f_*, respectively, as $n \to \infty$ uniformly over any compact subset of $(0, 1)^d$. This paper is the first to establish the consistency of the penalized convex regression estimator and its subgradient, thereby legitimizing the penalized convex regression estimator as an estimator of f_*.

In Section 2, we summarize notation and definitions. In Section 3, we describe our main results rigorously. We prove

of the main results in Section 4. We observe the numerical performance of the penalized convex regression estimator in Section 5. Section 6 includes some concluding remarks.

2. Notation and Definitions

For $z \in \mathbb{R}^d$, z^T denotes its transpose.

We say $f : (0,1)^d \to \mathbb{R}$ is differentiable at $x \in (0,1)^d$ if there is a vector $\nabla f(x) \in \mathbb{R}^d$ satisfying

$$\frac{f(x) - f(y) - \nabla f(x)^T (x - y)}{\|x - y\|} \to 0$$

as $y \to x$. We call $\nabla f(x)$ the gradient of f at x.

For a convex function $f : (0,1)^d \to \mathbb{R}$, we call $\xi \in \mathbb{R}^d$ a subgradient of f at $x \in (0,1)^d$ if $\xi^T(y - x) \leq f(y) - f(x)$ for all $y \in (0,1)^d$. We call the set of all subgradients at x the subdifferential at x, and denote it by $\partial f(x)$.

3. Main Results

To define the penalized convex regression estimator more rigorously, we first need to establish the existence of the solution to (3). Proposition 1 asserts that the solution to (3) exists uniquely.

Proposition 1 *The solution to (3) exists uniquely.*

Proof. (3) is a minimization problem of a continuous and coersive function over a non–empty closed domain, so the solution to (3) exists due to Proposition 7.3.1 and Theorem 7.3.7 on pp 216–217 of Kurdila and Zabarankin (2005). The solution is unique because the objective function of (3) is strictly convex. □

It should be noted that (3) computes the penalized convex regression estimator only at the X_i's. More specifically, if $(\hat{f}_1, \cdots \hat{f}_n, \hat{\xi}_1, \cdots, \hat{\xi}_n)$ is the solution to (3), then $\hat{f}_i$ is the penalized convex regression estimator computed at X_i. To define the penalized convex regression estimator at $x \neq X_i$, we set

$$\hat{g}_n(x) = \max_{1 \leq i \leq n} \{\hat{f}_i + \hat{\xi}_i^T (x - X_i)\} \tag{4}$$

for $x \in (0,1)^d$. The penalized convex regression estimator $\hat{g}_n$, defined by (4), will be the subject of study in this paper. In order to establish the consistency of $\hat{g}_n$ and its subgradient, we need to make the following assumptions.

A1. $\lambda_n \geq 0$ for all n and $\lambda_n \to 0$ as $n \to \infty$.

A2. $X_1, X_2, \cdots$ are iid $(0,1)^d$–valued random vectors.

A3. Given $X_1, X_2, \cdots, \varepsilon_1, \varepsilon_2, \cdots$ are iid random variables with a mean of zero and a finite variance.

A4. $f_* : (0,1)^d \to \mathbb{R}$ is convex and $\mathbb{E}[\sup_{\xi \in \partial f_*(X_1)} \|\xi\|^2] < \infty$.

A5. f_* is differentiable over $(0,1)^d$.

Our results are presented below.

Theorem 1 *Assume A1–A4. For any $\delta > 0$,*

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

with probability one as $n \to \infty$.

Theorem 2 *Assume A1–A5. For any $\delta > 0$,*

$$\sup\{\|\xi - \nabla f_*(x)\| : \xi \in \partial \hat{g}_n(x), x \in [\delta, 1 - \delta]^d\} \to 0$$

with probability one as $n \to \infty$.

4. Proofs of the Main Results

In this section, we prove Theorems 1 and 2.

4.1 Proof of Theorem 1

The proof consists of 10 steps.

Step 1: Since f_* is convex,

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{g}_n(X_i)-Y_i)^2+\frac{\lambda_n}{n}\sum_{i=1}^{n}\|\nabla\hat{g}_n(X_i)\|^2\leq\frac{1}{n}\sum_{i=1}^{n}(f_*(X_i)-Y_i)^2+\frac{\lambda_n}{n}\sum_{i=1}^{n}\|\nabla f_*(X_i)\|^2,$$

which implies

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}(\hat{g}_n(X_i)-f_*(X_i))^2\\ &\leq\ \frac{2}{n}\sum_{i=1}^{n}\varepsilon_i(\hat{g}_n(X_i)-f_*(X_i))+\frac{\lambda_n}{n}\sum_{i=1}^{n}\|\nabla f_*(X_i)\|^2-\frac{\lambda_n}{n}\sum_{i=1}^{n}\|\nabla\hat{g}_n(X_i)\|^2\\ &\leq\ \frac{2}{n}\sum_{i=1}^{n}\varepsilon_i(\hat{g}_n(X_i)-f_*(X_i))+\frac{\lambda_n}{n}\sum_{i=1}^{n}\|\nabla f_*(X_i)\|^2.\end{aligned}\tag{5}$$

Step 2: We next prove that

$$\frac{1}{n}\sum_{i=1}^{n}\hat{g}_n(X_i)^2\leq C\tag{6}$$

for some constant C and n sufficiently large a.s.

To see why this is true, let $\theta:(0,1)^d\to\mathbb{R}$ be defined by $\theta(x)=0$ for any $x\in(0,1)^d$. Then

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{g}_n(X_i)-Y_i)^2+\frac{\lambda_n}{n}\sum_{i=1}^{n}\|\nabla\hat{g}_n(X_i)\|^2\leq\frac{1}{n}\sum_{i=1}^{n}(\theta(X_i)-Y_i)^2+\frac{\lambda_n}{n}\sum_{i=1}^{n}\|\nabla\theta(X_i)\|^2.\tag{7}$$

Since $\nabla\theta(x)=(0,\cdots,0)\in\mathbb{R}^d$ for all $x\in(0,1)^d$, (7) implies

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{g}_n(X_i)-Y_i)^2\leq\frac{1}{n}\sum_{i=1}^{n}Y_i^2.\tag{8}$$

Thus, (8) and the Cauchy-Schwarz inequality imply

$$\frac{1}{n}\sum_{i=1}^{n}\hat{g}_n(X_i)^2\leq\frac{2}{n}\sum_{i=1}^{n}Y_i\hat{g}_n(X_i)\leq 2\sqrt{\frac{1}{n}\sum_{i=1}^{n}Y_i^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\hat{g}_n(X_i)^2},$$

and hence,

$$\frac{1}{n}\sum_{i=1}^{n}\hat{g}_n(X_i)^2\leq\frac{4}{n}\sum_{i=1}^{n}Y_i^2.$$

By the strong law of large numbers (SLLN),

$$\frac{1}{n}\sum_{i=1}^{n}\hat{g}_n(X_i)^2\leq 4\mathbb{E}[f_*(X_1)^2]+1$$

for n sufficiently large a.s., proving (6).

Step 3: We use Step 2 and the SLLN to show that for any subset of $(0,1)^d$, say A, with a nonempty interior, there exists a constant C_A satisfying

$$\inf_{x\in A}|\hat{g}_n(x)-f_*(x)|\leq C_A\tag{9}$$

a.s. for n sufficiently large.

To establish (9), suppose, on the contrary, we have

$$\inf_{x\in A}|\hat{g}_n(x)-f_*(x)|>\sqrt{2(C+\mathbb{E}[f_*(X_1)^2]+1)/\mathbb{P}(X_1\in A)}$$

with a positive probability. This implies

$$\begin{aligned}
&\liminf_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\sum_{i=1}^{n}(\hat{g}_n(X_i)-f_*(X_i))^2\\
&\geq \liminf_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}(\hat{g}_n(X_i)-f_*(X_i))^2 I(X_i\in A)\\
&\geq \liminf_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}I(X_i\in A)\liminf_{n\to\infty}\frac{\sum_{i=1}^{n}(\hat{g}_n(X_i)-f_*(X_i))^2 I(X_i\in A)}{\sum_{i=1}^{n}I(X_i\in A)}\\
&\geq \mathbb{P}(X_1\in A)2(C+\mathbb{E}[f_*(X_1)^2]+1)/\mathbb{P}(X_1\in A)\\
&= 2(C+\mathbb{E}[f_*(X_1)^2]+1) \qquad (10)
\end{aligned}$$

with a positive probability. (10) contradicts the fact that

$$\liminf_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}(\hat{g}_n(X_i)-f_*(X_i))^2\leq\liminf_{n\to\infty}\frac{2}{n}\sum_{i=1}^{n}\hat{g}_n(X_i)^2+\liminf_{n\to\infty}\frac{2}{n}\sum_{i=1}^{n}f_*(X_i)^2\leq 2(C+\mathbb{E}[f_*(X_1)^2]$$

a.s. for n sufficiently large by (6) and the SLLN. Hence, (9) follows.

Step 4: We use Step 2, Step 3, and the convexity of $\hat{g}_n$ to show that for any $\delta>0$, there is a constant C_δ satisfying

$$\sup_{x\in[\delta,1-\delta]^d}|\hat{g}_n(x)|\leq C_\delta \qquad (11)$$

a.s. for n sufficiently large. (11) follows from similar arguments to the proofs of Lemmas 3.2 and 3.3 on page 1644 of Seijo and Sen (2011).

Step 5: We note that (11) and Robert and Varberg (1974) imply that, for any $\delta>0$, there is a constant $\tilde{C}_\delta$ satisfying

$$|\hat{g}_n(x)-\hat{g}_n(y)|\leq\tilde{C}_\delta\|x-y\| \qquad (12)$$

for all $x,y\in[\delta,1-\delta]^d$ and n sufficiently large a.s.

Step 6: For $\delta>0$, let

$$\mathcal{F}_\delta=\{f:[\delta,1-\delta]^d\to\mathbb{R}\text{ such that }f\text{ is convex},\ |f(x)|\leq C_\delta,\ |f(x)-f(y)|\leq\tilde{C}_\delta\|x-y\|\text{ for all }x,y\in[\delta,1-\delta]^d\}.$$

By Steps 4 and 5, $\hat{g}_n$ restricted to $[\delta,1-\delta]^d$ belongs to $\mathcal{F}_\delta$ for n sufficiently large a.s.

For any $\epsilon>0$, there is a finite number of functions $f_1,\cdots,f_r$ in $\mathcal{F}_\delta$ with $r=r(\epsilon)$ satisfying, for any $f\in\mathcal{F}_\delta$,

$$\sup_{x\in[\delta,1-\delta]^d}|f(x)-f_i(f)|<\epsilon \qquad (13)$$

for some $i\in\{1,\cdots,r\}$; see, for example, Theorem 6 of Bronshtein (1976).

Step 7: We will utilize (6) and (13) to show that

$$\limsup_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(\hat{g}_n(X_i)-f_*(X_i))\leq 0\quad\text{a.s.} \qquad (14)$$

To fill in the details, let $\epsilon>0$ be given and set $A_\delta=[\delta,1-\delta]^d$ and $B_\delta=(0,1)^d\setminus[\delta,1-\delta]^d$. Note that

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(\hat{g}_n(X_i)-f_*(X_i))\\
&= \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(\hat{g}_n(X_i)-f_*(X_i))I(X_i\in A_\delta)+\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(\hat{g}_n(X_i)-f_*(X_i))I(X_i\in B_\delta)\\
&= I+II,\text{ say.}
\end{aligned}$$

By the Cauchy–Schwarz inequality and (6), we have

$$\begin{aligned} II &\leq \sqrt{\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2 I(X_i \in B_\delta)}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{g}_n(X_i)-f_*(X_i))^2} \\ &\leq \sqrt{\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2 I(X_i \in B_\delta)}\sqrt{\frac{2}{n}\sum_{i=1}^{n}\hat{g}_n(X_i)^2+\frac{2}{n}\sum_{i=1}^{n}f_*(X_i)^2} \\ &\leq \sqrt{\mathbb{E}[\varepsilon_1^2 I(X_1 \in B_\delta)]}\sqrt{2C+2\mathbb{E}[f_*(X_1)^2]} \end{aligned} \tag{15}$$

for n sufficiently large a.s. By taking δ small enough so that (15) $\leq 2\epsilon$, we can ensure

$$II < 2\epsilon \tag{16}$$

for n sufficiently large a.s.

On the other hand, we note that

$$\begin{aligned} I &= \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(\hat{g}_n(X_i)-f_*(X_i))I(X_i \in A_\delta) \\ &= \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(\hat{g}_n(X_i)-f_j(X_i))I(X_i \in A_\delta)+\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(f_j(X_i)-f_*(X_i))I(X_i \in A_\delta) \end{aligned}$$

for any $j \in \{1,\cdots,r(\epsilon)\}$, and hence,

$$I \leq \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(\hat{g}_n(X_i)-f_j(X_i))I(X_i \in A_\delta)+\max_{1\leq j\leq r(\epsilon)}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(f_j(X_i)-f_*(X_i))I(X_i \in A_\delta)$$

for any $j \in \{1,\cdots,r(\epsilon)\}$.

Therefore,

$$\begin{aligned} I &\leq \min_{1\leq j\leq r(\epsilon)}\frac{1}{n}\sum_{i=1}^{n}|\varepsilon_i||\hat{g}_n(X_i)-f_j(X_i)|I(X_i \in A_\delta)+\max_{1\leq j\leq r(\epsilon)}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(f_j(X_i)-f_*(X_i))I(X_i \in A_\delta) \\ &\leq \sqrt{\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2 I(X_i \in A_\delta)}\min_{1\leq j\leq r(\epsilon)}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{g}_n(X_i)-f_j(X_i))^2}+\max_{1\leq j\leq r(\epsilon)}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(f_j(X_i)-f_*(X_i))I(X_i \in A_\delta) \\ &\leq \epsilon\sqrt{\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2 I(X_i \in A_\delta)}+\max_{1\leq j\leq r(\epsilon)}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(f_j(X_i)-f_*(X_i))I(X_i \in A_\delta) \\ &\leq \epsilon(\mathbb{E}[\varepsilon_1^2]+1)+\epsilon \end{aligned} \tag{17}$$

for n sufficiently large a.s. by the SLLN.

By (16) and (17),

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

proving (14).

Step 9: By combining A4, (5) and (14), we obtain

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{g}_n(X_i)-f_*(X_i))^2 \to 0 \tag{18}$$

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

Step 10: We use (18) and Step 5 to establish Theorem 1. Let $\epsilon > 0$ be given. We divide $[\delta, 1-\delta]^d$ into compact subsets $S_1,\cdots,S_m$ with non-empty interior such that $\cup_{i=1}^{m}S_i \supset [\delta,1-\delta]^d$ and $\sup\{\|x-y\| : x,y \in S_i\} \leq \epsilon$ for $1 \leq i \leq m$.

By Step 5 and the fact that f_* is uniformly Lipschitz over $[\delta, 1-\delta]^d$, there is a constant $\tilde{C}$ satisfying $\|\hat{g}_n(x)-\hat{g}_n(y)\| \le \tilde{C}\|x-y\|$ and $|f_*(x) - f_*(y)| \le \tilde{C}\|x - y\|$ for all $x, y \in [\delta, 1 - \delta]^d$.

For any $x, X_j \in S_i$,

$$\begin{aligned} |\hat{g}_n(x) - f_*(x)| &\le |\hat{g}_n(x) - \hat{g}_n(X_j)| + |\hat{g}_n(X_j) - f_*(X_j)| + |f_*(X_j) - f_*(x)| \\ &\le \|x - X_j\|\tilde{C} + |\hat{g}_n(X_j) - f_*(X_j)| + \|x - X_j\|\tilde{C} \\ &\le 2\epsilon\tilde{C} + |\hat{g}_n(X_j) - f_*(X_j)|. \end{aligned}$$

Hence, for any $i \in \{1, \cdots, m\}$ and $x \in S_i$,

$$|\hat{g}_n(x) - f_*(x)| \le 2\epsilon\tilde{C} + \min_{X_j \in S_i} |\hat{g}_n(X_j) - f_*(X_j)|$$

and

$$\begin{aligned} \sup_{x \in S_i} |\hat{g}_n(x) - f_*(x)| &\le 2\epsilon\tilde{C} + \frac{1}{\sum_{j=1}^n I(X_j \in S_i)} \sum_{j=1}^n |\hat{g}_n(X_j) - f_*(X_j)| I(X_j \in S_i) \\ &\le 2\epsilon\tilde{C} + \frac{n}{\sum_{j=1}^n I(X_j \in S_i)} \frac{1}{n} \sum_{j=1}^n |\hat{g}_n(X_j) - f_*(X_j)| I(X_j \in S_i) \\ &\le 2\epsilon\tilde{C} + \frac{n}{\sum_{j=1}^n I(X_j \in S_i)} \sqrt{\frac{1}{n} \sum_{j=1}^n (\hat{g}_n(X_j) - f_*(X_j))^2} \\ &\le \epsilon C' \text{ by (18)} \end{aligned} \tag{19}$$

a.s. for n sufficiently large for some constant C'. Since (19) holds for each $i \in \{1, \cdots, m\}$, Theorem 1 follows.

4.2 Proof of Theorem 2

Suppose there is $\epsilon > 0$ and $x_1, x_2, \cdots \in [\delta, 1 - \delta]^d$ such that

$$\|\xi - \nabla f_*(x_n)\| \ge \epsilon \tag{20}$$

for $\xi \in \partial\hat{g}_n(x_n)$ and infinitely many n with a positive probability.

By (20), there is $i \in \{1, \cdots, d\}$ satisfying

$$|e_i^T \xi - e_i^T \nabla f_*(x_n)| \ge \epsilon/d \tag{21}$$

for $\xi \in \partial\hat{g}_n(x_n)$ and infinitely many n with a positive probability, where $e_i \in \mathbb{R}^d$ is a vector of zeros except for 1 in the ith entry for $1 \le i \le d$.

(21) implies either

$$e_i^T \xi \ge e_i^T \nabla f_*(x_n) + \epsilon/d, \tag{22}$$

or

$$e_i^T \nabla f_*(x_n) - \epsilon/d \ge e_i^T \xi \tag{23}$$

for infinitely many n with a positive probability. Suppose (22) holds. Then, there exists a subsequence $x_{n_1}, x_{n_2}, \cdots$ such that $x_{n_k} \to x_0 \in [\delta, 1 - \delta]^d$ as $k \to \infty$.

Note that by the definition of the subgradient, for any $h > 0$,

$$e_i^T \xi \le \frac{\hat{g}_{n_k}(x_{n_k} + he_i) - \hat{g}_{n_k}(x_{n_k})}{h}.$$

By the continuity of $\hat{g}_n$ and Theorem 1, letting $k \to \infty$ in both sides yields

$$e_i^T \xi \le \frac{f_*(x_0 + he_i) - f_*(x_0)}{h}. \tag{24}$$

By (22) and (24), we have

$$\limsup_{k \to \infty} e_i^T \nabla f_*(x_{n_k}) + \epsilon/d \le \frac{f_*(x_0 + he_i) - f_*(x_0)}{h}$$

and by the continuity of $\nabla f_*(\cdot)$, we have

$$e_i^T \nabla f_*(x_0) + \epsilon/d \leq \frac{f_*(x_0 + he_i) - f_*(x_0)}{h} \tag{25}$$

for any $h > 0$.

By letting $h \downarrow 0$ in (25), we obtain

$$e_i^T \nabla f_*(x_0) + \epsilon/d \leq e_i^T \nabla f_*(x_0),$$

which is a contradiction. Similarly, when we assume (23) holds, we can reach a contradiction. Hence, Theorem 2 is proved.

5. Empirical Studies

In this section, we compare the numerical behavior of the penalized convex regression estimator to that of the convex regression estimator. In Section 5.1, we assume that $f_* : (0, 1)^3 \to \mathbb{R}$ is given by a mathematical formula. In Section 5.2, $f_* : (1.2, 1.7) \to \mathbb{R}$ is the long run average waiting time per customer in an M/M/1 queue. Our findings from the numerical experiments are summarized in Section 5.3.

5.1 Simplified Example

We assume that $f_* : (0, 1)^3 \to \mathbb{R}$ is given by $f_*(z_1, z_2, z_3) = z_1^2 + 0.5z_2^2 + 0.2z_3^2$ for $(z_1, z_2, z_3) \in (0, 1)^3$, The X_i's are drawn uniformly from $(0, 1)^3$. The Y_i's are drawn from $Y_i = f_*(X_i) + \varepsilon_i$, where the ε_i's follow the normal distribution with mean 0 and standard deviation 0.1. Once the (X_i, Y_i)'s are obtained, we computed the convex regression estimator by solving (2) using CVX (Grant and Boyd (2014)). We also computed the penalized regression estimator by solving (3) with $\lambda_n = 1/n$ and $\lambda = 1/(2n)$, respectively, using CVX. To evaluate the performance of the penalized regression estimator $\hat{g}_n(\cdot)$, we computed the integrated mean square error (IMSE) between f_* and the estimator as follows:

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

and the IMSE between the gradient of f_* and the subgradient of $\hat{g}_n$ as follows:

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{\xi}_i - \nabla f_*(X_i))^2,$$

where the $\hat{\xi}_i$'s are the subgradients of $\hat{g}_n$ at the X_i's, computed from (2).

We then repeated this procedure 400 times independently, generating 400 IMSE values between f_* and $\hat{g}_n$ and 400 IMSE values between ∇f_* and the subgradient of $\hat{g}_n$. Using these values, we computed the 95% confidence interval of the IMSE between f_* and $\hat{g}_n$ and the 95% confidence interval of the IMSE between the gradient of f_* and the subgradient of $\hat{g}_n$. The IMSE values between f_* and the convex regression estimator is computed similarly. We reported the 95% confidence intervals between f_* and the estimators in Table 1 for a wide rage of n. We also report the 95% confidence intervals between the gradient of f_* and the subgrdient of the estimators in Table 2 for a wide range of n.

Table 1. The 95% confidence intervals of the IMSE between f_* and the estimators when $f_*(z_1, z_2, z_3) = z_1^2 + 0.5z_2^2 + 0.2z_3^2$ for $(z_1, z_2, z_3) \in (0, 1)^3$

n	Convex Regression Estimator	Penalized Convex Regression Estimator with $\lambda_n = 1/n$	Penalized Convex Regression Estimator with $\lambda_n = 1/(2n)$
20	0.0073 ± 0.0003	0.0091 ± 0.0003	0.0064 ± 0.0002
40	0.0053 ± 0.0002	0.0061 ± 0.0002	0.0034 ± 0.0001
60	0.0046 ± 0.0001	0.0040 ± 0.0001	0.0028 ± 0.0001

5.2 Single Server Queue

We assume that $f_*(x)$ is the long run average waiting time per customer in an M/M/1 queue, where the service times follow the exponential distribution with mean $1/x$ for $x \in (1.2, 1.7)$, and the interarrival times follow the exponential distribution with mean 1.

Table 2. The 95% confidence intervals of the IMSE values between the gradient of f_* and the subgradient of the estimators when $f_*(z_1, z_2, z_3) = z_1^2 + 0.5z_2^2 + 0.2z_3^2$ for $(z_1, z_2, z_3) \in (0, 1)^3$

n	Convex Regression Estimator	Penalized Convex Regression Estimator with $\lambda_n = 1/n$	Penalized Convex Regression Estimator with $\lambda_n = 1/(2n)$
20	$2.73 \times 10^8 \pm 1.56 \times 10^8$	0.1945 ± 0.0019	0.1554 ± 0.0018
40	$1.04 \times 10^{15} \pm 0.58 \times 10^{15}$	0.1274 ± 0.0014	0.0917 ± 0.0013
60	$8.43 \times 10^{17} \pm 1.97 \times 10^{17}$	0.0984 ± 0.0010	0.0768 ± 0.0010

The X_i's are drawn uniformly from $(1.2, 1.7)$. For each X_i, Y_i is generated by averaging the waiting times of the first 5000 customers in the single server queue, initialized empty and idle, with the service rate of X_i. Once the (X_i, Y_i)'s are obtained, we computed the convex regression estimator by solving (2) using CVX. We also computed the penalized regression estimator by solving (3) with $\lambda_n = 1/(20n)$ and $\lambda = 1/(40n)$, respectively, using CVX.

We reported the 95% confidence intervals between f_* and the estimators, using 400 iid trials, in Table 3 for a wide range of n. We also reported the 95% confidence intervals between the gradient of f_* and the subgrdients of the estimators, using 400 iid trials, in Table 4 for a wide range of n.

Table 3. The 95% confidence intervals of the IMSE between f_* and the estimators when f_* is the long-run average waiting time per customer in a M/M/1 queue

n	Convex Regression Estimator	Penalized Convex Regression Estimator with $\lambda_n = 1/(20n)$	Penalized Convex Regression Estimator with $\lambda_n = 1/(40n)$
10	0.0850 ± 0.0103	0.0373 ± 0.0028	0.0719 ± 0.0058
20	0.0436 ± 0.0058	0.0267 ± 0.0018	0.0322 ± 0.0025
30	0.0262 ± 0.0048	0.0214 ± 0.0014	0.0172 ± 0.0016

Table 4. The 95% confidence intervals of the IMSE between the gradient of f_* and the subgradient of the estimators whenf_* is the long run average waiting time per customer in an M/M/1 queue

n	Convex Regression Estimator	Penalized Convex Regression Estimator with $\lambda_n = 1/(20n)$	Penalized Convex Regression Estimator with $\lambda_n = 1/(40n)$
10	$3.50 \times 10^{10} \pm 3.48 \times 10^{10}$	16.63 ± 0.40	56.53 ± 1.33
20	$1.47 \times 10^{16} \pm 2.35 \times 10^{16}$	12.90 ± 0.42	16.43 ± 0.82
30	$1.61 \times 10^{17} \pm 1.14 \times 10^{17}$	11.21 ± 0.48	9.34 ± 0.55

5.3 Observations from Numerical Experiments

Tables 1 and 3 show that both the convex regression estimator and the penalized convex regression estimator converge in terms of the IMSE. On the other hand, Tables 2 and 4 indicate that only the subgradient of the penalized convex regression estimator shows convergence in terms of the IMSE, and that the subgradient of the convex regression estimator diverges as n increases in terms of the IMSE. This phenomenon is consistent with what we observed in Figure 1; the convex regression estimator has extremely large subgradients near the boundary of the domain, while the penalized convex regression estimator has bounded subgradients throughout the domain.

It is also observed that the numerical performance of the penalized convex regression estimator is highly dependent on the smoothing constant λ_n. Thus, the issue of how to select the smoothing constant is a promising future research topic.

6. Conclusions

In this paper, we established consistency of the penalized convex regression estimator. Numerical experiments show that, unlike the convex regression estimator, the penalized convex regression estimator and its gradient are convergent near the boundary of its domain. Hence, a promising research topic for the future is a thorough examination of the behavior of the penalized convex regression estimator near the boundary of its domain.

References

Bertsimas, D., & Mundru, N. (2020). Sparse convex regression. *INFORMS J. Comput.* https://doi.org/10.1287/ijoc.2020.0954

Boyd, S., & Vandenberghe, L. (2004). *Convex Optimization.* Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/CBO9780511804441

Bronshtein, E. M. (1976). ϵ-entropy of convex sets and functions. *Siberian Math. J., 17,* 393–398. https://doi.org/10.1007/BF00967858

Chen, X., Lin, Q., & Sen, B. (2020). On degrees of freedom of projection estimators with applications to multivariate nonparametric regression. *J. Amer. Statist. Assoc., 115,* 173–186. https://doi.org/10.1080/01621459.2018.1537917

Grant, M., & Boyd, S. (2014). CVX: Matlab software for disciplined convex programming, version 2.1. Retrieved May 2020, from http://cvxr.com/cvx

Groeneboom, P., Jongbloed, G., & Wellner, J. A. (2001). Estimation of a convex function: Characterizations and asymptotic theory. *Ann. Statist., 29,* 1653–1698. https://doi.org/10.1214/aos/1015345958

Hanson, D. L., & Pledger, G. (1976). Consistency in concave regression. *Ann. Statist., 4,* 1038–1050. https://doi.org/10.1214/aos/1176343640

Hildreth, C. (1954). Point Estimates of Ordinates of Concave Functions. *J. Amer. Statist. Assoc., 49*, 598–619. https://doi.org/10.1080/01621459.1954.10483523

Kuosmanen, T. (2008). Representation theorem for convex nonparametric least squares. *Econometrics J., 11,* 308–325. https://doi.org/10.1111/j.1368-423X.2008.00239.x

Kurdila, A. J., & Zabarankin, M. (2005). *Convex functional analysis (Systems & control: Foundations & applications)*. Switzerland: Birkhäuser.

Mazumder, R., Choudhury, A., Iyengar, G., & Sen, B. (2019). Computational framework for multivariate convex regression and its variants. *J. Amer. Statist. Assoc., 114* 318–331. https://doi.org/10.1080/01621459.2017.1407771

Roberts, A. W., & Varberg, D. E. (1974). Another proof that convex functions are locally Lipschitz. *Amer. Math. Monthly, 81,* 1014–1016. https://doi.org/10.1080/00029890.1974.11993721

Seijo, E., & Sen, B. (2011). Nonparametric least squares estimation of a multivariate convex regression function. *Ann. Statist., 39,* 1633–1657. https://doi.org/10.1214/10-AOS852

Varian, H. R. (1984). The nonparametric approach to production analysis. *Econometrica, 52,* 579–597. https://doi.org/10.2307/1913466

Weber, R. R. (1983). A note on waiting times in single server queues. *Oper. Res., 31,* 950–951. https://doi.org/10.1287/opre.31.5.950

Copyrights