On the distance between Cumulative Sum Diagram and its Gretest Convex Minorant for unequally spaced design points

JAYANTA KUMAR PAL
University of Michigan
MICHAEL WOODROOFE
University of Michigan.

Running headline : Cusum diagram and its convex minorant.

Abstract

The supremum difference between the cumulative sum diagram, and its greatest convex minorant, in case of nonparametric isotonic regression is considered. When the regression function is strictly increasing, and the design points are unequally spaced, but approximate a positive density in even a slow rate ($n^{-1/3}$), then the difference is shown to shrink in a very rapid (close to $n^{-2/3}$) rate. The result is analogous to the corresponding result in case of a monotone density estimation established by Kiefer and Wolfowitz, but uses entirely different representation. The limit distribution of the GCM as a process on the unit interval is obtained when the design variables are i.i.d. with a positive density. Finally, a pointwise asymptotic normality result is proved for the smooth monotone estimator, obtained by the convolution of a kernel with the classical monotone estimator.

Keywords : greatest convex minorant, Hungarian representation, Marshall's Lemma, modulus of continuity, monotone function.

1 Introduction

A common problem of the nonparametric regression theory is to estimate a monotone increasing (or decreasing) function. Examples include the estimation of growth curves, survival functions, reliability functions, renewal functions and biomedical problems like dose-response and epidemic of disease. The maximum likelihood estimate of those functions is obtained by computing the greatest convex minorant $\tilde{G}_n$ of the cumulative sum diagram G_n of the available data. This is cumbersome to represent and less tractable to manage theoretically. However, the computation of the constrained MLE of the underlying regression function cannot circumnavigate that essential part, since it is precisely the left hand slope of the greatest convex minorant (GCM now on) by definition.

Though the properties and the large sample behaviour of the regression function have been fairly well-studied in literature, the GCM itself is also of relevance in some cases like survival functions, reliability functions, increasing return-to-scale items in consumer behavior etc., as it estimates the cumulative regression function. The unconstrained estimator being G_n, it follows from Marshall's lemma, that $\|\tilde{G_n} - G_0\| \leq \|G_n - G_0\|$, where G_0 is the true cumulative regression function, and $\|.\|$ is the supremum norm. That is, $\tilde{G}_n$ is a more accurate estimator of G_0 than G_n. This article shows that the unconstrained version G_n is not very different from $\tilde{G}_n$, however.

Notation is established in Section 2, and some preliminary results are obtained. In Section 3, we prove, under modest conditions on the distribution of the regressors and the underlying monotone functions,

$$\|G_n - \tilde{G_n}\| = O_P(n^{-2/3}(\log n)^{2/3}). \tag{1}$$

That renders $\tilde{G}_n$ as efficient as G_n asymptotically to estimate G_0. However, $\tilde{G}_n$ is convex by construction, whereas G_n may not have that property. For the case of equally spaced regressors (1) was obtained by Durot & Tocquet(2003). Kiefer & Wolfowitz(1976) established a similar result for the distance between the empirical distribution function and its least concave majorant under monotonicity constraints on the density of the observations. Recently, Kulikov & Lopuhaä(2004) investigated a properly scaled version of that difference, and proved that this difference converges as a process in distribution to the corresponding process for two-sided Brownian motion with parabolic

drift. Here we concentrate on the regression scenario and relax the conditions on the regressors. Our extension is important because it allows the regressors to be realizations of i.i.d. random variables, provided these have a bounded positive density.

In Section 4 we describe some applications of the main result. To begin we study the asymptotic properties of the GCM as a process. Since the deviance is negligible compared to the usual rate of convergence($\sqrt{n}$) for the empirical process, the GCM process also has a Gaussian large-sample distribution. The limit is the sum of a time transformation of a Brownian Motion and an independent integral with respect to a Brownian Bridge. As a corollary to that convergence in law, we establish a central limit theorem for functionals of the form $\int_0^t \hat{\mu}(t)f(t)dt$ where f is of bounded variation and $\hat{\mu}$ is the NPMLE for the regression function. The functional processes converge faster than the estimate $\hat{\mu}$, and the limit process is Gaussian, provided that f has a bounded derivative. The resulting estimator has a limit distribution in the form of a stochastic integral process, with a $\sqrt{n}$ rate. The asymptotic normality is also established in a case where f changes with the sample size. The kernel smoothed version of the isotonic estimator is the function under consideration. Clearly, the kernel changes with varying bandwidth b_n, and as the latter goes to zero, the kernel converges to an infinite spike, which is no longer of bounded variation. Even in this case, if the bandwidth does not shrink too rapidly (at least slower than $n^{-1/3}$), then the corresponding estimator is shown to be asymptotically normal, with the usual rate $\sqrt{nb_n}$.

The problem is investigated only in case of an increasing regression function (i.e. convex cumulative function). Consequently, the entire theory can be easily translated in case of a decreasing (concave resp.) function, after simply redefining the problem and using a backward orientation.

As we discuss the supremum distance between the CSD and the GCM in this article, it is worth mentioning that the corresponding pointwise distance has been investigated in Meyer & Woodroofe(2000). They actually show that if $m/n \to x_0$, where n is the sample size, then $n^{2/3}(G_n(m) - \tilde{G}_n(m))$ converges in distribution to a non-degenerate distribution. However, since we are investigating the supremum of those pointwise distances, the convergence rate is marginally slower in our case.

2 Preliminary results

The underlying regression function μ is assumed to be continuously differentiable on the closed unit interval. For simplicity and without loss of generality, we restrict our domain to be $[0,1]$. Further restrictions on the shape of μ will be imposed shortly.

To estimate μ, one selects a set of regressors $\{t_i, i = 1, \ldots, n\}$ in increasing order. The empirical distribution function of them over $[0,1]$ is defined as

$$\Lambda_n(t) = \frac{1}{n}\#\{i : t_i \leq t\}$$

The regressors may even be the order statistics of an i.i.d. sequence. We prefer to keep the dependence of that sequence on n implicit, by not calling them $t_{i,n}$, only to avoid complication. In general, they form a triangular array of reals in increasing order.

The data comes in the form of $Y_1, Y_2, \ldots, Y_n$, observations taken at those points with μ as the mean function and i.i.d. errors. i.e.

$$Y_i = \mu(t_i) + \epsilon_i$$

where ϵ_i's are conditionally i.i.d. given the regressors from a distribution F with zero mean and a finite moment generating function. The cumulative sum diagram (CSD) G_n is the unique piecewise linear function with knots at $\{1, \ldots, n\}$ defined on the interval $[0, n]$ and values

$$G_n(j) = \frac{1}{n}\sum_{i=1}^{j} Y_i$$

at the knots.

$\tilde{G}_n$ is the largest convex function H defined on $[0, n]$ for which $H \leq G_n$. We seek to estimate the maximum distance between G_n and $\tilde{G}_n$ and find how fast it converges to 0.

We will show that, under the assumptions A1-A3, we can bound the differences uniformly with rate close to $n^{-2/3}$.

Assumptions:

A1 There exists a distribution function Λ with density λ on $[0,1]$ such that,

$$\pi_n = \sup_t |\Lambda_n(t) - \Lambda(t)| = o(n^{-1/3}).$$

A2 There exists constants a and b such that $0 < a \leq \lambda(t) \leq b$ for all t.

A3 There exists constants A and B such that,

$$0 < A \leq \mu'(t) \leq B \text{ for all } t.$$

In particular, A1 encompasses the case of equally spaced design variables as well as i.i.d. regressors from a distribution with bounded positive density. In the former case, $\Lambda(t) = t$, and $\pi_n = 1/n$. In the second case, A1 follows a.s. from the law of the Iterated Logarithm for empirical distribution functions. A conditioning on those regressors yield the final result.

The assumption A2 is also fairly non-restrictive since it allows all the continuous non-vanishing densities with a bounded support including uniform, truncated normal, truncated exponential etc. However, the densities which either vanish or blow up near the endpoints do not satisfy the assumption. In general, the beta densities fall in that category. Assumption A3 merely requires the regression function to be increasing in a stricter sense. Now, we are in a position to state the main result of this article in a succinct form.

Theorem 1 *Under assumptions A1, A2, A3, (1) holds.*

As the rigorous proof of the statement requires a few coherent important results, we proceed towards it in a logical sequence.

We start defining the partial sums of errors $S_i = \epsilon_1 + \ldots + \epsilon_i$. In order to measure the fluctuations in that sequence, we have to approximate that by their limit process. Convergence in law (as given by the Donsker's principle) will not be enough as we need to measure the error in that approximation as well. The Hungarian representation of the partial sums, as proved by Komlos, Major & Tusnady(1975), gives us the desired approximation. We formally state the result as follows.

Hungarian Representation :

Let $\epsilon_1, \epsilon_2, \ldots$ be i.i.d. random variables with finite moment generating functions in a neighborhood of the origin. Then its partial sums $\{S_n\}$ satisfy the following. There exists a probability space Ω, a sequence of random variables $\{T_n\}$ and a Brownian motion $\{\mathcal{B}(t) : 0 \leq t \leq 1\}$, both embedded in the aforementioned probability space whose existence is ascertained, such that,

1. $\{S_n\}$ and $\{T_n\}$ has the same distribution as sequences.

2. We define $R_n(t) = T_{\lfloor nt \rfloor} - \sqrt{n}\mathcal{B}(t)$. Then,

$$\sup_{0 \le t \le 1} |R_n(t)| = O(\log n) \text{ a.s.}$$

Now we will consider a subset of the regressor points, placed at strategic distances and consider the linear interpolation of G_n at those intermediate knots. The goal is to show that the resulting process is convex with probability approaching 1 as n gets large.

For every n, we choose an integer k_n. Further restrictions will be imposed on k_n, later in the argument. Let $l_n = \lceil n/k_n \rceil$, and $m_n = l_n k_n$. It is clear that $m_n/n \to 1$, as long as l_n also goes to infinity as n grows. We observe that, $\|G_n - \tilde{G}_n\| \le \|G_{m_n} - \tilde{G}_{m_n}\|$, by noting that the GCM of G_{m_n} restricted to $[0, n]$ is a convex function lying below G_n, and hence falls below $\tilde{G}_n$ as well. So, it is enough to prove (1) for sample size m_n, as

$$\frac{m_n^{-\frac{2}{3}} (\log m_n)^{\frac{2}{3}}}{n^{-\frac{2}{3}} (\log n)^{\frac{2}{3}}} \to 1.$$

Hence, without loss of generality, we will assume that k_n divides n and $l_n = n/k_n$. The new set of knots is $\{s_i, i = 1, \ldots, k_n\}$ where $s_i = t_{il_n}$ such that $\Lambda_n(s_i) = i/k_n$. Let G^n be the continuous piecewise linear function with knots at $\{1, \ldots, n\}$ and values

$$G^n(j) = \frac{1}{n} \sum_{i=1}^{j} \mu(t_i) \text{ for } j = 1, \ldots, n$$

at the knots. Since, μ is an increasing function, G^n is convex for each n. Next, let

$$\begin{aligned} L_n^{k_n}(s) &= \frac{s - (i-1)l_n}{l_n} G_n(il_n) + \frac{il_n - s}{l_n} G_n((i-1)l_n) \\ \text{and } L^{k_n}(s) &= \frac{s - (i-1)l_n}{l_n} G^n(il_n) + \frac{il_n - s}{l_n} G^n((i-1)l_n) \end{aligned}$$

for $(i-1)l_n \le s < il_n$ and $i = 1, \ldots, k_n$. For typographical simplicity, we write k, l, G, L_n, L in place of $k_n, l_n, G^n, L_n^{k_n}, L^{k_n}$ respectively, where the dependence on n and k_n is understood. Therefore, G is the mean function of the corresponding process, whereas L_n and L serves as the interpolated versions of G_n and G respectively, with knots spaced at $il, i = 1, \ldots, k$ and joined

linearly in between. Figure 1 (in the Appendix) illustrates all those functions for small n, namely $n = 12, k = 3, l = 4$. Let $A_n = \{L_n \text{ is convex}\}$. Then

$$\begin{aligned} A_n^c &= \cup_{i=2}^{k-1}\{G_n((i+1)l) + G_n((i-1)l) < 2G_n(il)\} \\ &= \cup_{i=2}^{k-1}\{n(G((i+1)l) + G((i-1)l) - 2G(il)) + S_{(i+1)l} + S_{(i-1)l} - 2S_{il} < 0\} \end{aligned}$$

Lemma 1 *If $k\pi_n \to 0$ then, there exists a_0 and b_0 positive constants such that,*

$$\frac{a_0}{k} \leq s_{i+1} - s_i \leq \frac{b_0}{k} \text{ for all } i.$$

Proof :

$$\begin{aligned} \frac{1}{k} = \Lambda_n(s_{i+1}) - \Lambda_n(s_i) &\leq \Lambda(s_{i+1}) - \Lambda(s_i) + 2\pi_n \\ &\leq b(s_{i+1} - s_i) + 2\pi_n \end{aligned}$$

So, $s_{i+1} - s_i \geq 1/bk - 2\pi_n/b \geq a_0/k$ for large n and some a_0. Similarly,

$$\begin{aligned} \frac{1}{k} = \Lambda_n(s_{i+1}) - \Lambda_n(s_i) &\geq \Lambda(s_{i+1}) - \Lambda(s_i) - 2\pi_n \\ &\geq a(s_{i+1} - s_i) - 2\pi_n \end{aligned}$$

So, $s_{i+1} - s_i \leq 1/ak + 2\pi_n/a \leq b_0/k$ for large n and some b_0. With a modification of a_0 and b_0, the inequalities extend for all n.

Proposition 1 *If $k\pi_n \to 0$ then, there is a positive β for which*

$$G((i+1)l) + G((i-1)l) - 2G(il) \geq \frac{\beta}{k^2} \tag{2}$$

for all $i = 2, \ldots, k-1$ and for all $n \geq 3$.

Proof : To start, we know that $\mu'(t) \geq A > 0$ for all values of t. Also, as $k\pi_n \to 0$, we can find a large enough n_0, and $\beta > 0$ for which

$$k\pi_n < \frac{Aaa_0^2 - \beta}{4Ab_0}$$

for $n > n_0$. Now we express the left side of (2) as an integral of the function μ'. For $n > n_0$,

$$
\begin{aligned}
& G((i+1)l) + G((i-1)l) - 2G(il) \\
= & \int_{s_i}^{s_{i+1}} \mu(r)\, d\Lambda_n(r) - \int_{s_{i-1}}^{s_i} \mu(r)\, d\Lambda_n(r) \\
= & \int_{s_i}^{s_{i+1}} \{\mu(s_i) + \int_{s_i}^{r} \mu'(s)ds\}\, d\Lambda_n(r) - \int_{s_{i-1}}^{s_i} \{\mu(s_i) - \int_{r}^{s_i} \mu'(s)ds\}\, d\Lambda_n(r) \\
\geq & \mu(s_i)\frac{1}{k} + A\int_{s_i}^{s_{i+1}} \int_{s_i}^{r} ds\, d\Lambda_n(r) - \mu(s_i)\frac{1}{k} + A\int_{s_{i-1}}^{s_i} \int_{r}^{s_i} ds\, d\Lambda_n(r) \\
= & A[\int_{s_i}^{s_{i+1}} (\Lambda_n(s_{i+1}) - \Lambda_n(s))ds + \int_{s_{i-1}}^{s_i} (\Lambda_n(s) - \Lambda_n(s_{i-1}))ds] \qquad (3) \\
\geq & A[\int_{s_i}^{s_{i+1}} a(s_{i+1} - s)ds - 2\pi_n(s_{i+1} - s_i) + \int_{s_{i-1}}^{s_i} a(s - s_{i-1})ds - 2\pi_n(s_i - s_{i-1})] \\
\geq & \frac{1}{2}Aa[(s_{i+1} - s_i)^2 + (s_i - s_{i-1})^2] - 2A\pi_n(s_{i+1} - s_{i-1}) \\
\geq & Aa\frac{a_0^2}{k^2} - 2A\pi_n\frac{2b_0}{k} \\
\geq & \frac{\beta}{k^2}.
\end{aligned}
$$

Also, (3) implies that, $G((i+1)l) + G((i-1)l) - 2G(il) > 0$ for all $k, n \geq 3$. Therefore, after modifying the definition of β for the finitely many low values of $3 \leq n \leq n_0$, we extend the inequality for all n.

Now we have to fix the intermediate distances in such a way that the linearized process is convex for large n almost surely.

Proposition 2 *Suppose there exists* $0 < \alpha < \beta^2/16$ *such that,* $(\alpha n/\log n)^{1/3} \leq k_n \leq (\beta^2 n/(16\log n))^{1/3}$, *where* β *is as in Proposition 1. Then,*

$$P(A_n^c) \to 0$$

as n *gets large i.e.* L_n *becomes a convex function with probability approaching 1.*

Proof : As $k\pi_n \to 0$, we conclude,

$$\begin{aligned} P(A_n^c) &\leq P[\min_{i=2}^{k-1}\{S_{(i+1)l} + S_{(i-1)l} - 2S_{il}\} < -\frac{n\beta}{k^2}] \\ &= P[\min_{i=2}^{k-1}\{T_{(i+1)l} + T_{(i-1)l} - 2T_{il}\} < -\frac{n\beta}{k^2}] \\ &\leq P[\min_{i=2}^{k-1}(\mathcal{B}(\frac{i+1}{k}) - 2\mathcal{B}(\frac{i}{k}) + \mathcal{B}(\frac{i-1}{k})) < -\frac{\beta l^2}{2n\sqrt{n}}] \\ &\quad + P[\min_{i=2}^{k-1}(R_n(\frac{i+1}{k}) - 2R_n(\frac{i}{k}) + R_n(\frac{i-1}{k})) < -\frac{\beta l^2}{2n}] \end{aligned}$$

As $l^2/(n\log n) \to \infty$, we get

$$\lim P[\min_{i=2}^{k-1}(R_n(\frac{i+1}{k}) - 2R_n(\frac{i}{k}) + R_n(\frac{i-1}{k})) < -\frac{\beta l^2}{2n}] = 0$$

from the strong approximation theorem, since $\sup_t |R_n(t)| = O(\log n/n)$ with probability 1. For the first term, we observe that for any $t > 0$, $\Phi(-t) \leq \phi(t)/t$, where Φ and ϕ denote the standard normal cumulative distribution function and the density function respectively. Then, since $l = n/k$,

$$\begin{aligned} & P[\min_{i=2}^{k-1}(\mathcal{B}(\frac{i+1}{k}) - 2\mathcal{B}(\frac{i}{k}) + \mathcal{B}(\frac{i-1}{k})) < -\frac{\beta l^2}{2n\sqrt{n}}] \\ \leq\ & k\Phi(-\frac{\beta\sqrt{n}}{2\sqrt{2k^3}}) \\ \leq\ & \frac{2k^2\sqrt{k}}{\beta\sqrt{\pi n}} e^{-\frac{\beta^2 n}{16k^3}} \\ \leq\ & \frac{2k^2\sqrt{k}}{\beta n\sqrt{\pi n}} \\ \to\ & 0 \end{aligned}$$

Hence, $P(A_n^c)$ decreases to 0 as n grows large. From now on, suppose that $(\alpha n/\log n)^{1/3} \leq k_n \leq (\beta^2 n/(16\log n))^{1/3}$.

3 CSD and the corresponding GCM

We now proceed to estimate the difference between the cumulative sum function and its greatest convex minorant using the new functions.

Lemma 2 *Under A_n,*

$$\|G_n - \tilde{G}_n\| \leq 2[\|G_n - L_n - G + L\| + \|L - G\|]$$

Proof: Marshalls' Lemma says that, for any convex h, $\|\tilde{G}_n - h\| \leq \|G_n - h\|$. In particular, under A_n, L_n satisfies the property. Therefore,

$$\begin{aligned} \|G_n - \tilde{G}_n\| &\leq \|G_n - L_n\| + \|L_n - \tilde{G}_n\| \\ &\leq 2\|G_n - L_n\| \\ &\leq 2[\|G_n - L_n - G + L\| + \|L - G\|]. \end{aligned}$$

That concludes the proof of the lemma.

The first term is actually the difference between a centered process and its interpolated version, which will be approximated using Levy's modulus of uniform continuity. However, the second term, the difference between the mean function and the linear interpolation, can be shown to be positive and sufficiently small using the convexity of the mean function and the boundedness of μ'. The idea is to represent the difference in term of an integral, the integrand being μ' itself. We want to remind the reader that $\Lambda_n(s_i) = i/k$.

Lemma 3 *For the given values of k_n, There exists $\Gamma > 0$ such that,*

$$0 \leq L(s) - G(s) \leq \frac{\Gamma}{k^2}$$

for all s. The constant Γ depends on the suprema of μ' as well as λ.

Proof : As G is convex, the non-negativity is immediate. However, to establish the upper bound, we need to express the difference in an integral form. We note that, as $k\pi_n \to 0$, we have $\tau = \sup k\pi_n < \infty$. We define, $\Gamma = Bbb_0^2 + 2Bb\tau$. Clearly, it is enough to consider only the values

at the knots. As $\mu' < B$, we have, for $(i-1)l \le j < il$,

$$
\begin{aligned}
& L(j) - G(j) \\
= \; & \frac{j-(i-1)l}{l}[G(il) - G(j)] - \frac{il-j}{l}[G(j) - G((i-1)l)] \\
= \; & \frac{j-(i-1)l}{l}\int_{t_j}^{t_{il}} \mu(r)d\Lambda_n(r) - \frac{il-j}{l}\int_{t_{(i-1)l}}^{t_j} \mu(r)d\Lambda_n(r) \\
= \; & \frac{j-(i-1)l}{l}\int_{t_j}^{t_{il}} \{\mu(t_j) + \int_{t_j}^{r} \mu'(s)ds\}d\Lambda_n(r) - \frac{il-j}{l}\int_{t_{(i-1)l}}^{t_j} \{\mu(t_j) - \int_{r}^{t_j} \mu'(s)ds\}d\Lambda_n(r) \\
\le \; & B[\frac{j-(i-1)l}{l}\int_{t_j}^{t_{il}} (\Lambda_n(t_{il}) - \Lambda_n(s))ds + \frac{il-j}{l}\int_{t_{(i-1)l}}^{t_j} (\Lambda_n(s) - \Lambda_n(t_{(i-1)l}))ds] \\
\le \; & B[\frac{j-(i-1)l}{l}\int_{t_j}^{t_{il}} (\Lambda(t_{il}) - \Lambda(s) + 2\pi_n)ds + \frac{il-j}{l}\int_{t_{(i-1)l}}^{t_j} (\Lambda(s) - \Lambda(t_{(i-1)l}) + 2\pi_n)ds] \\
\le \; & B[\frac{j-(i-1)l}{l}\{b\int_{t_j}^{t_{il}} (t_{il} - s)ds + 2\pi_n(t_{il} - t_j)\} \\
& + \frac{il-j}{l}\{b\int_{t_{(i-1)l}}^{t_j} (s - t_{(i-1)l})ds + 2\pi_n(t_j - t_{(i-1)l})\}] \\
= \; & Bb[\frac{j-(i-1)l}{2l}(t_{il} - t_j)^2 + \frac{il-j}{2l}(t_j - t_{(i-1)l})^2] \\
& + 2B\pi_n[\frac{j-(i-1)l}{l}(t_{il} - t_j) + \frac{il-j}{l}(t_j - t_{(i-1)l})] \\
\le \; & Bb\frac{b_0^2}{k^2} + 2B\pi_n\frac{b_0}{k} \\
\le \; & \frac{\Gamma}{k^2}.
\end{aligned}
$$

Hence, the proof is complete.

Now we proceed to estimate the difference between the centered process and its interpolated version. We use the Hungarian representation again, to bound the deviances using Levy's modulus of continuity.

Proposition 3 *For any value of k,*

$$\|G_n - G - L_n + L\| = O_P(\sqrt{\frac{\log k}{nk}}) + O_P(\frac{\log n}{n}).$$

Proof : We remind the reader that, there exists T_n, with same distribution as S_n, and $\sup|T_{\lfloor nt\rfloor} - \sqrt{n}\mathcal{B}(t)| = \sup|R_n(t)| = O(\log n/n)$ almost surely. The modulus of uniform continuity can be stated as,

$$\limsup_{\delta\downarrow 0}\frac{\max_{|t-s|<\delta}|\mathcal{B}(t)-\mathcal{B}(s)|}{\sqrt{2\delta\log\frac{1}{\delta}}} = 1$$

a.s. See Karatzas & Shreve (2000) for a rigorous proof. Moreover, we observe that, since all the involved functions are piecewise linear, the supremum actually occurs at one of the knots. Hence, we can write,

$$\begin{aligned}
& \sup_t|G_n(t)-G(t)-L_n(t)+L(t)| \\
= \;& \frac{1}{n}\max_{i=1,\ldots,k}\max_{(i-1)l\le j<il}|S_j - \frac{j-(i-1)l}{l}S_{il} - \frac{il-j}{l}S_{(i-1)l}| \\
\stackrel{D}{=} \;& \frac{1}{n}\max_{i=1,\ldots,k}\max_{(i-1)l\le j<il}|T_j - \frac{j-(i-1)l}{l}T_{il} - \frac{il-j}{l}T_{(i-1)l}| \\
= \;& \frac{1}{n}\max_i\max_{(i-1)l\le j<il}|\sqrt{n}\{\mathcal{B}(\frac{j}{n}) - \frac{j-(i-1)l}{l}\mathcal{B}(\frac{i}{k}) - \frac{il-j}{l}\mathcal{B}(\frac{i-1}{k})\} \\
& + R_n(\frac{j}{n}) - \frac{j-(i-1)l}{l}R_n(\frac{i}{k}) - \frac{il-j}{l}R_n(\frac{i-1}{k})| \\
\le \;& \frac{1}{\sqrt{n}}\sup_{0\le t-s\le\frac{1}{k}}|\mathcal{B}(s)-\mathcal{B}(t)| + \frac{2}{n}\sup|R_n(t)| \\
= \;& O_P(\sqrt{\frac{\log k}{nk}}) + O_P(\frac{\log n}{n}).
\end{aligned}$$

Note that all the inequalities hold almost surely in the arguments, using Levy's modulus of uniform continuity and the strong approximation. However, since we are only working with an identical copy of the actual variables, our conclusion reaches only stochastic boundedness, rather than boundedness with probability 1.

Proof of Theorem 1

Let k be as selected as in Proposition 2. Then, using Proposition 3 we get,

$$\begin{aligned} \|G_n - L_n - G + L\| &= O_P\big(\sqrt{\frac{\log k}{nk}}\big) + O_P\big(\frac{\log n}{n}\big) \\ &= O_P\big((\frac{n}{\log n})^{-\frac{2}{3}}\big) \qquad \text{for our choice of } k. \end{aligned}$$

Also, for those k, from Lemma 3,

$$\|L - G\| = O\big((\frac{n}{\log n})^{-\frac{2}{3}}\big)$$

The theorem can be proved now as follows. Under A_n, both the terms $\|L-G\|$ and $\|G_n-L_n-G+L\|$ are stochastically bounded with magnitude $(n/\log n)^{-2/3}$. In addition, the complement set A_n^c vanishes in probability as n gets larger. As a consequence, (1) is automatic.

4 Applications

To start with a simple example, suppose we have equally spaced design variables in a bounded subset of the real line. With a linear transformation, we can work with $t_i = i/n$. In this case, $\pi_n = 1/n$, and $\Lambda(t) = t$ for t in $[0, 1]$. A simple application of the result gives the rate of shrinking. As shown by Durot & Tocquet(2003), the magnitude is precisely of order $n^{-2/3}(\log n)^{2/3}$, and not less than that.

To consider the other extreme, let the regressors be chosen randomly from a distribution Λ with strictly positive density λ and then sorted. The errors are independent of the regressors as well. Then, we deduce from the Law of the Iterated Logarithm, that

$$\limsup \sqrt{\frac{n}{2\log\log n}}\|\Lambda_n - \Lambda\|_\infty \leq \frac{1}{2} \text{ a.s.}$$

For a proof, see Van der Vaart(1998). As $n^{1/3}\|\Lambda_n - \Lambda\|_\infty \to 0$ almost surely, we can restrict our process on the set $A = \{n^{1/3}\|\Lambda_n - \Lambda\|_\infty \to 0\}$, which has probability 1. Clearly, $\|G_n - \tilde{G}_n\| = O_P(n/\log n)^{-2/3}$ here as well, with arguments similar to the proof of the main theorem, after conditioning on the regressors $t_i, i = 1, \ldots, n$.

Distribution of the GCM as a process :

We start defining the process $H_n(t) = \frac{1}{n}\sum_{t_i \le t} Y_i$, the underlying empirical process. Let $\tilde{H}_n(t)$ be the process defined by the GCM of the corresponding CSD, as defined before. This is not the GCM of the empirical process itself, and so, not necessarily a convex function on the unit interval. To establish its asymptotic distribution, we need to consider the empirical process of the regressors to get an estimate of the difference, since the means of the process contribute to the final variance. As a consequence of the theorem, and using the fluctuations of a Brownian motion, we establish the following result.

Proposition 4 *Under the above definitions, if* $\{T_1, T_2, \ldots, T_n\}$ *are i.i.d. regressors from a distribution* Λ *with positive bounded density* λ*, and the errors are independent of those regressors, then*

$$\sqrt{n}(\tilde{H}_n(t) - \int_0^t \mu(s)\lambda(s)ds) \Rightarrow \Gamma(t)$$

where $\Gamma(t) : t \in [0,1]$ *is defined as the sum of an integral with respect to a Brownian Bridge* $\mathcal{W}$ *and a time-transformed Brownian motion* $\mathcal{B} \circ \Lambda$*.*

Proof: Define $\bar{H}_n(t)$ as the piecewise linear function with knot values $\bar{H}_n(T_{(i)}) = H_n(T_{(i)})$. Now,

$$H_n(t) - \int_0^t \mu(s)\lambda(s)ds = \int_0^t \mu(s)[d\Lambda_n(s) - d\Lambda(s)] + \frac{1}{n}S_{n\Lambda_n(t)}$$

The first term, if multiplied by $\sqrt{n}$, converges to $\int_0^t \mu(s)d\mathcal{W}(s)$ in distribution. To see this, we observe that the function $R \mapsto \int_0^t \mu(s)dR(s)$ is a continuous function from $\mathcal{D}[0,1]$ onto itself, by virtue of the boundedness of μ'. Since the empirical process $\sqrt{n}[\Lambda_n - \Lambda]$ converges to a Brownian Bridge $\mathcal{W}$, the result follows.

As mentioned in Hungarian embedding, the sequence $\{S_i\}$ can be shown to be identically distributed as another sequence of random variables, which can be approximated by a Brownian motion, with proper rate of convergence. For simplicity, we denote the new sequence as S_i as well, since it does not alter the approximation. Hence, we can approximate it uniformly and almost surely, with an

error of magnitude $(\log n/n)$ as,

$$\begin{aligned} & \frac{1}{\sqrt{n}}\mathcal{B}(\Lambda_n(t)) \\ = \quad & \frac{1}{\sqrt{n}}\mathcal{B}(\Lambda(t)) + \frac{1}{\sqrt{n}}(\mathcal{B}(\Lambda_n(t) - \mathcal{B}(\Lambda(t)))) \end{aligned}$$

The second term can be bounded uniformly in probability by $\frac{1}{\sqrt{n}}\sqrt{\|\Lambda_n - \Lambda\| \log \frac{1}{\|\Lambda_n - \Lambda\|}}$, which is of smaller magnitude than $\frac{1}{\sqrt{n}}$. This bound is obtained using Levy's modulus of uniform continuity. Combining these bounds, we deduce, $\sqrt{n}(H_n(t) - \int_0^t \mu(s)\lambda(s)ds) \Rightarrow \int_0^t \mu(s)d\mathcal{W}(s) + \mathcal{B}(\Lambda(t))$. Moreover, as the observational errors are independent of the regressors, so are the two terms in the above expression.

To conclude the proof, we observe,

$$\begin{aligned} \sup_t |H_n(t) - \tilde{H}_n(t)| \quad &\leq \quad \sup_t |H_n(t) - \bar{H}_n(t)| + \sup_t |\bar{H}_n(t) - \tilde{H}_n(t)| \\ &= \quad \max_i |\frac{Y_i}{n}| + \|G_n - \tilde{G}_n\| \\ &\leq \quad \max_i \frac{|\epsilon_i|}{n} + o_p(\frac{1}{n}) + O_p(n^{-\frac{2}{3}}(\log n)^{2/3}) \\ &= \quad o_p(\frac{1}{\sqrt{n}}) \end{aligned}$$

Hence, the proposition follows.

From now on, define the function $H(t) = \int_0^t \mu(s)\lambda(s)ds$. Let $\hat{\mu}$ denote the NPMLE of μ under the monotonicity constraint, and recall that this is the left hand slope of the GCM $\tilde{G}_n$ itself. It has been well-established that $\hat{\mu}$ is a consistent estimator, albeit in a slow $(n^{-1/3})$ rate. For example, see Robertson, Wright & Dykstra(1987). Its limit distribution is different from Gaussian, given by a non-linear functional of a Brownian Motion.

Integral functionals :

Instead of directly estimating μ, sometimes scientists might be interested in estimating what can be called an integral functional of μ, namely $\int_0^t f(s)\mu(s)ds = \tilde{\Psi}[f](t)$, for some pre-specified f, which is a weighted average of the regression function. An obvious estimator will be

$1/n \sum_{t_i \leq t} f(t_i)\hat{\mu}(t_i) = \hat{\Psi}[f](t)$. However, because of the non-uniform sampling scheme, that actually estimates $\int_0^t f(s)\mu(s)\lambda(s)ds = \Psi[f](t)$. Our next proposition investigates the estimator and its large sample behavior.

Proposition 5 *Under the same set of assumptions, for every continuously differentiable function f with a bounded derivative,*

$$\sqrt{n}(\hat{\Psi}[f](t) - \Psi[f](t)) \Rightarrow \int_0^t f(s)d\Gamma(s)$$

as a process on the unit interval. The limit process is an integral in the Ito sense, as the integrating variable is a random process.

Proof: We start by noting that $\hat{\Psi}[f](t) = \int_0^t f(s)\hat{\mu}(s)d\Lambda_n(s)$ and $\tilde{H}_n(t) = \int_0^t \hat{\mu}(s)d\Lambda_n(s)$. Using the differentiability of f, we can write,

$$\begin{aligned}
& \sqrt{n}(\hat{\Psi}[f](t) - \Psi[f](t)) \\
= \; & \sqrt{n}[f(t)\tilde{H}_n(t) - \int_0^t f'(s)\tilde{H}_n(s)ds - f(t)H(t) + \int_0^t f'(s)H(s)ds] \\
= \; & f(t)[\sqrt{n}\{\tilde{H}_n(t) - H(t)\}] - \int_0^t f'(s)[\sqrt{n}\{\tilde{H}_n(s) - H(s)\}]ds
\end{aligned}$$

The functional $\mathcal{U} : \mathcal{C}[0,1] \to \mathcal{C}[0,1]$ defined by $\mathcal{U}[R](t) = f(t)R(t) - \int_0^t f'(s)R(s)ds$ is continuous in the supremum topology, since f' is a bounded function. Hence, as a corollary to Proposition 4, and using the continuous mapping theorem, we conclude,

$$\begin{aligned}
\sqrt{n}(\hat{\Psi}[f](t) - \Psi[f](t)) \; & \Rightarrow \; f(t)\Gamma(t) - \int_0^t f'(s)\Gamma(s)ds \\
& = \; \int_0^t f(s)d\Gamma(s)
\end{aligned}$$

Thus, we get the necessary convergence.

Kernel smoothing of the isotonic estimator :

In general, even if f changes with sample size, and does not have a bounded derivative, we can still derive the distribution of $\int_0^1 f(t)\hat{\mu}(t)dt$, provided it does not fluctuate too much. We discuss an application of this scenario.

As observed by Mukherjee(1988), the isotonic estimator $\hat{\mu}$ lacks smoothness. His modified estimator was defined using a kernel k and a bandwidth sequence $\{b_n\}$ converging to 0. To avoid complications, we write b only. Invoking his estimator, we define,

$$\tilde{\mu}(t) = \frac{\sum_{i=1}^{n} k(\frac{t-t_i}{b})\hat{\mu}(t_i)}{\sum_{i=1}^{n} k(\frac{t-t_i}{b})} = \frac{\int_0^1 k(\frac{t-s}{b})\frac{1}{b}d\tilde{H}_n(s)}{\int_0^1 k(\frac{t-s}{b})\frac{1}{b}d\Lambda_n(s)}$$

Defining $H_n(t)$ as $1/n\sum_{t_i \leq t} Y_i$, we can establish the asymptotic normality for the estimator. The following proposition proves that phenomenon under modest conditions on the bandwidth sequence.

Proposition 6 *Suppose the sequence of bandwidths satisfies the constraint* $nb^3/(\log n)^4 \to \infty$, *but* $nb^5 \to 0$. *Also suppose the kernel* k *is symmetric, bounded and has a bounded support on* $\mathbb{R}$. *The design variables are assumed to come from a common distribution with a positive density* λ *on* $[0,1]$. *Then, for* $0 < t < 1$, *we have,*

$$\sqrt{nb}(\tilde{\mu}(t) - \mu(t)) \Rightarrow N[0, \frac{\sigma^2}{\lambda(t)} \int k^2(u)du]$$

Proof: We observe that $\|H_n - \tilde{H}_n\| = O_P(n/\log n)^{-2/3}$, using Theorem 1. For that estimate, we deduce,

$$\begin{aligned}
& |\int_0^1 \frac{1}{b}k(\frac{t-s}{b})d\tilde{H}_n(s) - \int_0^1 \frac{1}{b}k(\frac{t-s}{b})dH_n(s)| \\
= \; & |\frac{1}{b}k(\frac{t-1}{b})[\tilde{H}_n(1) - H_n(1)] + \int_0^1 \frac{1}{b^2}k'(\frac{t-s}{b})[\tilde{H}_n(s) - H_n(s)]ds| \\
\leq \; & \|H_n - \tilde{H}_n\|\big[\frac{1}{b}k(\frac{t-1}{b}) + \int_0^1 \frac{1}{b^2}|k'(\frac{t-s}{b})|ds\big] \\
= \; & O_P\big(\frac{1}{b}n^{-\frac{2}{3}}(\log n)^{\frac{2}{3}}\big) \qquad\qquad \text{since } \int |k'(u)|du < \infty \\
= \; & o_P(\frac{1}{\sqrt{nb}})
\end{aligned}$$

Now, we recall from the theory of kernel estimation the following two results.

$$\sqrt{nb}\big[\frac{\sum_i k(\frac{t-t_i}{b})Y_i}{\sum_i k(\frac{t-t_i}{b})} - \mu(t)\big] \Rightarrow N[0, \frac{\sigma^2}{\lambda(t)} \int k^2(u)du] \tag{4}$$

$$\frac{1}{nb}\sum_i k(\frac{t-t_i}{b}) = \lambda(t) + o_P(1) \tag{5}$$

Combining these two results with the above bounds, the proposition follows.
Since the isotonic smooth estimator has a normal limit which depends only on the variance σ^2 and the underlying density, confidence intervals can be constructed replacing them by their consistent estimators.

5 Conclusion

As we have established the asymptotic equivalence of G_n and $\tilde{G}_n$ in the monotone regression framework, the result can be viewed as a parallel to the corresponding result for density estimation established by Kiefer & Wolfowitz(1976). However, the basic nature of density estimation and regression are quite different, and this requires different treatment. As opposed to the large deviation theory and bounds on tail probabilities, we used strong approximation results to get hold of the partial sums by their limit process. However, the main theme running through this article is the proximity of the CSD and the GCM, which relates to the aforementioned article.
The result yielding the asymptotic normality of the GCM process will serve as an important consequence. It will be useful to construct confidence intervals of the true cumulative regression function, if we know the random distribution for the regressors. We can as well use $\bar{H}_n$ or $\tilde{H}_n$ as the mean of the confidence band, where the width shrinks with a $\sqrt{n}$ rate. Regarding the evaluation functionals, it should be also noted that the regression function can be characterized by its evaluation functionals over a class of functions, and hence a study of their behavior will be of substantial importance.
Finally, as shown in the case of monotone smooth estimator, kernel estimation and isotonic estimation can be hybridized without disturbing the limit behavior of the estimators. More challenging problems involving the cumulative regression function can be addressed using the CSD and the GCM, by interchanging the role of one with the other, since they are asymptotically equivalent upto a fast rate.

Acknowledgements:
The research was supported by the National Science Foundation and a grant from the Horace Rackham Graduate School.

References

[1] Durot, C. & Tocquet, A.S. (2003). On the distance between the empirical process and its concave majorant in a monotone regression framework. Ann. Inst. H. Poincaré Probab. Statist. PR 39, 2, 217-240.

[2] Karatzas, I. & Shreve, S. (2000). Brownian Motion and Stochastic Calculus. Springer - Graduate Text in Mathematics.

[3] Kiefer, J. & Wolfowitz, J. (1976). Asymptotically minimax estimation of concave and convex distribution functions. Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete 34, 73-85.

[4] Komlos, J., Major, P. & Tusnady, G. (1975). An Approximation of Partial Sums of Independent RV'-s and the Sample DF. I. Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete 32,111-131.

[5] Kulikov, V.N. & Lopuhaä, H.P. (2004). The limit process of the difference between the empirical distribution function and its concave majorant. Preprint available on http://ssor.twi.tudelft.nl/∼lopuhaa.

[6] Meyer, M. & Woodroofe, M. (2000). On the Degrees of Freedom in Shape-restricted Regression. Ann. Statist. 28 1083-1104

[7] Mukherjee, H. (1998). Monotone Nonparametric Regression. Ann. Statist. 16 741-750

[8] Robertson, T., Wright, F. & Dykstra, R. (1987). Order restricted statistical inference. John Wiley and sons.

[9] Van der Vaart, A. (1998). Asymptotic Statistics. Cambridge Series in Statistical and Probabilistic Mathematics.

Jayanta Kumar Pal, Department of Statistics, University of Michigan, 1085 South University, 436 West Hall, Ann Arbor, MI 48109. U.S.A. (jpal@umich.edu)

Appendix: The plot of the functions defined in Section 2

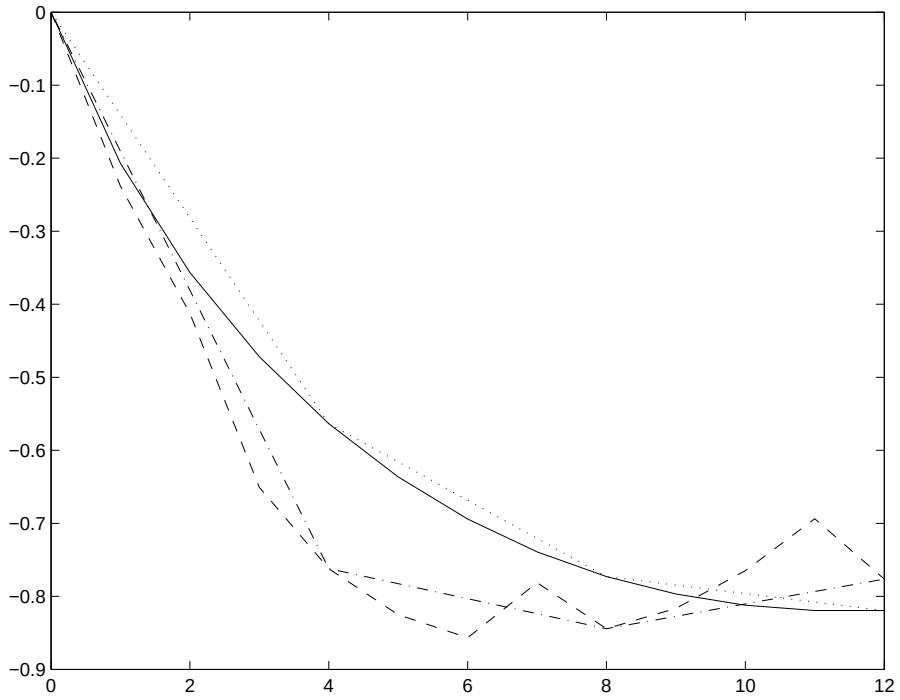

Fig. 1: We illustrate the functions G, G_n, L and L_n for $n = 12, k = 3, l = 4$. Those are represented by the solid line, the dashed line, the dotted line and the dash-dotted line respectively.