The Annals of Statistics
2021, Vol. 49, No. 4, 2021–2052
https://doi.org/10.1214/20-AOS2025

CONFIDENCE INTERVALS FOR MULTIPLE ISOTONIC REGRESSION AND OTHER MONOTONE MODELS

BY HANG DENG[*], QIYANG HAN[†] AND CUN-HUI ZHANG[‡]

Department of Statistics, Rutgers University, [*]hdeng@stat.rutgers.edu; [†]qh85@stat.rutgers.edu; [‡]czhang@stat.rutgers.edu

We consider the problem of constructing pointwise confidence intervals in the multiple isotonic regression model. Recently, Han and Zhang (2020) obtained a pointwise limit distribution theory for the so-called block max–min and min–max estimators (Fokianos, Leucht and Neumann (2020); Deng and Zhang (2020)) in this model, but inference remains a difficult problem due to the nuisance parameter in the limit distribution that involves multiple unknown partial derivatives of the true regression function.

In this paper, we show that this difficult nuisance parameter can be effectively eliminated by taking advantage of information beyond point estimates in the block max–min and min–max estimators. Formally, let $\widehat{u}(x_0)$ (resp. $\widehat{v}(x_0)$) be the maximizing lower-left (resp. minimizing upper-right) vertex in the block max–min (resp. min–max) estimator, and $\widehat{f}_n$ be the average of the block max–min and min–max estimators. If all (first-order) partial derivatives of f_0 are nonvanishing at x_0, then the following pivotal limit distribution theory holds:

$$\sqrt{n_{\widehat{u},\widehat{v}}(x_0)}\big(\widehat{f}_n(x_0) - f_0(x_0)\big) \rightsquigarrow \sigma \cdot \mathbb{L}_{\mathbf{1}_d}.$$

Here $n_{\widehat{u},\widehat{v}}(x_0)$ is the number of design points in the block $[\widehat{u}(x_0), \widehat{v}(x_0)]$, σ is the standard deviation of the errors, and $\mathbb{L}_{\mathbf{1}_d}$ is a universal limit distribution free of nuisance parameters. This immediately yields confidence intervals for $f_0(x_0)$ with asymptotically exact confidence level and oracle length. Notably, the construction of the confidence intervals, even new in the univariate setting, requires no more efforts than performing an isotonic regression once using the block max–min and min–max estimators, and can be easily adapted to other common monotone models including, for example, (i) monotone density estimation, (ii) interval censoring model with current status data, (iii) counting process model with panel count data, and (iv) generalized linear models. Extensive simulations are carried out to support our theory.

1. Introduction.

1.1. *Overview.* The field of estimation and inference under shape constraints has undergone rapid development in recent years, mostly notably in the direction of estimation theory of multidimensional shape constrained models. We briefly give some review of the history and some recent progress:

- (*Univariate shape constraints*) Starting from the seminal work of [25, 55, 56], estimation of a univariate monotone density or regression function has received much attention, cf. [8–10, 26, 27, 29, 38, 67, 69]. Estimation of a univariate convex density or regression function is a more challenging task, but considerable progress has been made through the efforts of many authors, cf. [8, 10, 32, 33, 35, 43, 44, 52]. Recent years also witnessed much progress in further understanding the behavior of the maximum likelihood estimator

Received January 2020; revised July 2020.
MSC2020 subject classifications. Primary 60E15; secondary 62G05.
Key words and phrases. Limit distribution theory, confidence interval, multiple isotonic regression, Gaussian process, shape constraints.

(MLE) of a univariate log-concave density, cf. [1, 17, 19, 20, 46, 47]. Other topics include estimation of unimodal regression functions, cf. [8, 12], estimation of a concave bathtub-shaped hazard function, cf. [45], and estimation of a k-monotone density, cf. [2].
- (*Multidimensional monotonicity constraints*) [11] initiated a study of risk bounds for the least squares estimator (LSE) of a bivariate coordinate-wise nondecreasing regression function. [39] extends the results of [11] to the more challenging case $d \geq 3$. See also [37] for some further improvements. [15] studied block max–min and min–max estimators originally proposed in [23]. See also a recent work [21] for a different notion of multidimensional monotonicity.
- (*Multidimensional convexity constraints*) Convex/concave regression in multidimensional settings is initiated in [50]. Consistency of the LSEs is proved in [51, 60]. [41] studied global and adaptive risk bounds for convex bounded LSEs in a random design for $d \leq 3$. [47] studied global risk bounds for log-concave MLEs, and [22] studied their adaptation properties, both for $d \leq 3$. [68] studied log-concave density estimation in high dimensions. Other topics include estimation of s-concave densities, cf. [37, 40, 48, 63], and additive modeling, cf. [13].

Despite these remarkable progress in the *estimation* theory of shape constrained models in multivariate settings for various tuning-free estimators, little next to nothing is known about how these merits can be actually useful in making *inference* for the multidimensional shape constrained function of interest. The purpose of this paper is to start to fill this gap, within the context of multiple isotonic regression [15, 39].

Here is our setup. Consider the regression model

$$Y_i = f_0(X_i) + \xi_i, \quad i = 1, \ldots, n, \tag{1.1}$$

where $X_1, \ldots, X_n$ are design points in $[0, 1]^d$ which can be either fixed or random, and $\xi_1, \ldots, \xi_n$ are independent mean-zero errors. The true regression function f_0 is assumed to belong to the class of coordinate-wise nondecreasing functions on $[0, 1]^d$:

$$f_0 \in \mathcal{F}_d \equiv \{f : [0, 1]^d \to \mathbb{R}, f(x) \leq f(y) \text{ if } x_i \leq y_i \text{ for all } i = 1, \ldots, d\}.$$

The hope of making some real progress in the inference aspect of this model, beyond purely the estimation theory, is spurred by the recent work of the second and third authors [42], who obtained a *pointwise limit distribution theory* for the block max–min and min–max estimators originally proposed in [23] and rigorously defined in [15]. For *any* $x_0 \in [0, 1]^d$, let the block max–min and min–max estimators, $\widehat{f}_n^-$ and $\widehat{f}_n^+$, be defined as

$$\begin{aligned}\widehat{f}_n^-(x_0) &\equiv \max_{u \leq x_0} \min_{\substack{v \geq x_0 \\ [u,v]\cap\{X_i\}\neq\varnothing}} \frac{1}{|\{i : u \leq X_i \leq v\}|} \sum_{i:u\leq X_i\leq v} Y_i \\ &\equiv \max_{u \leq x_0} \min_{\substack{v \geq x_0 \\ [u,v]\cap\{X_i\}\neq\varnothing}} \bar{Y}|_{[u,v]} \quad \text{and} \\ \widehat{f}_n^+(x_0) &\equiv \min_{v \geq x_0} \max_{\substack{u \leq x_0 \\ [u,v]\cap\{X_i\}\neq\varnothing}} \bar{Y}|_{[u,v]}.\end{aligned} \tag{1.2}$$

Note that in the univariate case ($d = 1$), the block max–min estimator $\widehat{f}_n^-$ and the block min–max estimator $\widehat{f}_n^+$ are the same and coincide with the isotonic least squares estimator (LSE) at design points $\{X_i\}$. However, $\widehat{f}_n^-$ and $\widehat{f}_n^+$ are in general different and $\widehat{f}_n^- \leq \widehat{f}_n^+$ is only guaranteed at design points in $d \geq 2$; see [15] for an explicit example in which the two estimators differ.

If the errors ξ_i's are i.i.d. mean-zero with variance σ^2, [42] showed that

$$\omega_n^{-1}(\boldsymbol{\alpha})\big(\widehat{f}_n^{\mp}(x_0) - f_0(x_0)\big) \rightsquigarrow r(\sigma) \cdot K(f_0, x_0) \cdot \mathbb{D}_{\boldsymbol{\alpha}}^{\mp}. \tag{1.3}$$

Here $\omega_n(\boldsymbol{\alpha})$ is the local rate of convergence of $\widehat{f}_n^{\mp}$, depending on the 'local smoothness' level $\boldsymbol{\alpha}$ of f_0 at x_0 (the precise meaning of this will be clarified in Section 2) and the design of the covariates in a fairly complicated way, $r(\sigma)$ is a constant depending on the noise level of the errors, and $K(f_0, x_0)$ is a constant depending on the unknown information concerning the derivatives of f_0 at x_0. [42] also showed that the limit distribution theory (1.3) is optimal in a local asymptotic minimax sense.

One may naturally wish to use (1.3) for construction of confidence intervals (CIs) for $f_0(x_0)$, but unfortunately, the complications for using directly the above limit theory for inference are multifold:

1. The constants $K(f_0, x_0), r(\sigma)$ depend on the unknown information of derivatives of f_0 at x_0 and the noise level σ;
2. The local rate of convergence $\omega_n^{-1}(\boldsymbol{\alpha})$ depends on the unknown local smoothness level $\boldsymbol{\alpha}$ of f_0.

Even one could be content with the knowledge of the local smoothness level of f_0 at x_0, for instance, assuming all first-order partial derivatives are nonvanishing, the problem of getting a consistent estimate of the nuisance parameter $K(f_0, x_0)$, which involves *many* derivatives of f_0 at x_0 is already very challenging. Since one of the main features of shape-constrained methods is the avoidance of tuning parameters—which is particularly important in multidimensional settings—we would ideally want to avoid estimation of derivatives to begin with.

A popular tuning-free testing approach for inference in the univariate monotone-response models, put forward in [3, 6], proposes the use of a log likelihood ratio test. The strength of this method lies in the fact that the limit distribution of the log likelihood ratio statistic is *pivotal*, that is, not depending on nuisance parameters, in particular the derivative of the monotone function of interest, provided it is nonvanishing at the point of interest. Using the quantiles for the pivotal limit distribution, one can then obtain CIs by inverting a family of log likelihood ratio tests. The same idea is further exploited in [16, 18] in the contexts of inference for the mode of a log-concave density and for the value of a concave regression function.

It is natural to wonder if a similar program, based on likelihood methods, can be extended to multidimensional settings, for instance in the multiple isotonic regression model (1.1) we study here. Apart from the apparent lack of any limit distribution theory for the LSE, that is, the maximum likelihood estimator under Gaussian likelihood, the more fundamental problem is that the LSE does exhibit some undesirable suboptimal behavior. In particular, as have been clear from the work [39], the LSE does not adapt to constant functions at the near optimal parametric rate, while the block max–min and min–max estimators (1.2) do [15, 42]. This strongly hints that a limit distribution theory of type (1.3) does not hold for the LSE, or at most can only hold for a very restrictive range of $\boldsymbol{\alpha}$, since (1.3) already recovers the parametric rate for constant signals.

Another common approach for avoiding estimation of nuisance parameters in limit distributions is the bootstrap. However, as shown in [49, 59, 62], standard bootstrap methods in nonstandard problems, in particular those with cube-root asymptotics and nonnormal limit distributions, typically lead to *inconsistent* estimates. Although it is in principle possible to develop consistent bootstrap procedures,for example, m-out-of-n bootstrap, or bootstrap with smoothing, cf. [59, 62], these procedures involve one or more tuning parameters that need to be carefully calibrated in practice, which unfortunately demerits the tuning-free advantages of shape-constrained methods.

The conceptual and practical difficulties in the likelihood and bootstrap methods lead us to a completely different approach for making inference of $f_0(x_0)$. Our proposal for the construction of the CI for $f_0(x_0)$, as will be detailed in (1.7) below, *requires essentially no*

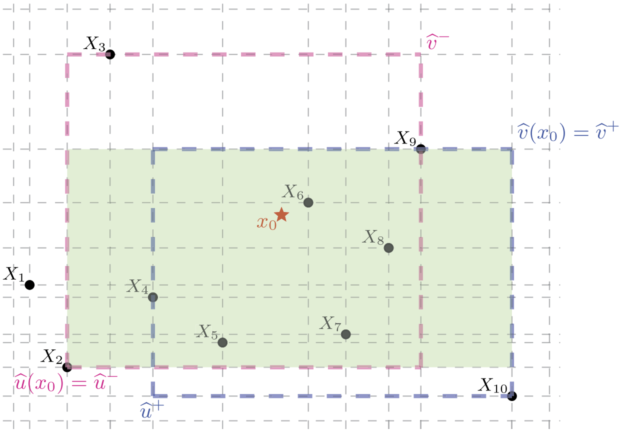


FIG. 1. *Figure illustration of the specification of* $\widehat{u}(x_0)$ *and* $\widehat{v}(x_0)$.

more efforts than performing an isotonic regression once using the block max–min and min–max estimators. The key idea for our proposal is to use information beyond point estimates in isotonic regression to directly estimate the scaled magnitude $\omega_n^* \equiv \omega_n(\boldsymbol{\alpha})r(\sigma)K(f_0, x_0)/\sigma$ of the error of estimating $f_0(x_0)$ in (1.3) and therefore to bypass the difficult problem of estimating the nuisance parameter $K(f_0, x_0)$. More important, the implementation of this idea does not require consistent estimation of the scaled magnitude: Given estimates $\widehat{f}_n(x_0)$ and $\widehat{\omega}_n^*$, it requires only the convergence in distribution of the product of $(\widehat{f}_n(x_0) - f_0(x_0)\}/\omega_n^*$ and the ratio $\omega_n^*/\widehat{\omega}_n^*$ to a known distribution. See Theorems 1 and 2 in the next section and their proofs.

Formally, let $(\widehat{u}(x_0), \widehat{v}(x_0))$ be any pair such that

$$
\begin{aligned}
\widehat{f}_n^-(x_0) &\equiv \max_{u \le x_0} \min_{\substack{v \ge x_0 \\ [u,v]\cap\{X_i\} \neq \varnothing}} \bar{Y}|_{[u,v]} = \min_{\substack{v \ge x_0 \\ [u,v]\cap\{X_i\} \neq \varnothing}} \bar{Y}|_{[\widehat{u}(x_0),v]}, \\
\widehat{f}_n^+(x_0) &\equiv \min_{v \ge x_0} \max_{\substack{u \le x_0 \\ [u,v]\cap\{X_i\} \neq \varnothing}} \bar{Y}|_{[u,v]} = \max_{\substack{u \le x_0 \\ [u,v]\cap\{X_i\} \neq \varnothing}} \bar{Y}|_{[u,\widehat{v}(x_0)]}.
\end{aligned}
\tag{1.4}
$$

Let the average of the two estimators $\widehat{f}_n^{\mp}$ in (1.2) be the *block average estimator* $\widehat{f}_n(x_0)$, that is,

$$
\widehat{f}_n(x_0) \equiv \frac{1}{2}\big(\widehat{f}_n^-(x_0) + \widehat{f}_n^+(x_0)\big),
\tag{1.5}
$$

and let $n_{\widehat{u},\widehat{v}}(x_0)$ be the number of design points in the block $[\widehat{u}(x_0), \widehat{v}(x_0)]$, that is,

$$
n_{\widehat{u},\widehat{v}}(x_0) \equiv \sum_i \mathbf{1}_{X_i \in [\widehat{u}(x_0),\widehat{v}(x_0)]}.
\tag{1.6}
$$

When $(Y_1, \dots, Y_n)$ are in general positions (i.e., $A \neq B$ implies $\bar{Y}|_A \neq \bar{Y}|_B$ for nonempty A and B), the set of design points in the rectangle $[\widehat{u}^-, \widehat{v}^-]$ giving $\widehat{f}_n^-(x_0)$ in (1.4) is unique. In this case, $\widehat{u}(x_0) = \widehat{u}^-$ is unique if we confine our choice to the rectangle $[\widehat{u}^-, \widehat{v}^-]$ with at least one design point in each of its 2^d sides. Similarly, $\widehat{v}(x_0)$ is also unique when the solution $[\widehat{u}^+, \widehat{v}^+]$ for $\widehat{f}_n^+(x_0)$ in (1.4) is required to have a design point on each side. For such specification of $(\widehat{u}(x_0), \widehat{v}(x_0))$, $n_{\widehat{u},\widehat{v}}(x_0)$ defined above is uniquely specified. See Figure 1 for an illustration. In any cases, our theoretical results hold for all feasible pairs $(\widehat{u}(x_0), \widehat{v}(x_0))$ in (1.4).

We propose the following form of CI:

$$
\mathcal{I}_n(c_\delta) \equiv \left[\widehat{f}_n(x_0) - \frac{c_\delta \cdot \widehat{\sigma}}{\sqrt{n_{\widehat{u},\widehat{v}}(x_0)}}, \widehat{f}_n(x_0) + \frac{c_\delta \cdot \widehat{\sigma}}{\sqrt{n_{\widehat{u},\widehat{v}}(x_0)}}\right],
\tag{1.7}
$$

where $\widehat{\sigma}$ is the square root of an estimator $\widehat{\sigma}^2$ of the variance σ^2, and $c_\delta > 0$ is a critical value chosen by the user that depends only on the confidence level $\delta > 0$.

The crux of our proposal (1.7) is the following *pivotal limit distribution theory*: Under the same conditions as in the limit distribution theory (1.3),

$$\sqrt{n_{\widehat{u},\widehat{v}}(x_0)}\big(\widehat{f}_n(x_0) - f_0(x_0)\big) \rightsquigarrow \sigma \cdot \mathbb{L}_{\boldsymbol{\alpha}}, \tag{1.8}$$

where the distribution of $\mathbb{L}_{\boldsymbol{\alpha}}$ *does not depend on the nuisance parameter* $K(f_0, x_0)$. Hence, given a consistent variance estimator $\widehat{\sigma}^2$, the CI (1.7) can both achieve asymptotically *exact* confidence level $1-\delta$ and shrink at the oracle length on the order of $\omega_n(\boldsymbol{\alpha}) \cdot r(\sigma) \cdot K(f_0, x_0)$, provided that the local smoothness $\boldsymbol{\alpha}$ is known. This is the case, for example, if one assumes that all first-order partial derivatives are nonvanishing, much as in [3, 6] in the univariate case that assumes a nonvanishing first derivative for the monotone function at the point of interest. By relaxing the requirement of asymptotically *exact* confidence level, it is also possible, by *calibrating the critical value c_δ alone*, to construct conservative CIs (1.7) that adapt to any given range of local smoothness levels $\boldsymbol{\alpha}$, while maintaining the optimal order of the length as in the limit theory (1.3). One natural question here is that whether the likelihood approach of [3, 6] in the much simpler univariate setting, wins over our proposal (1.7) in terms of adaptation to unknown local smoothness $\boldsymbol{\alpha}$. As will be clear in Section 2, the limit distributions of log likelihood ratio tests also depend on $\boldsymbol{\alpha}$, so indeed the likelihood approach of [3, 6] by itself does not offer a stronger degree of universality from the perspective of adaptation.

At a deeper level, the viewpoint of our construction for the CI for $f_0(x_0)$ is markedly different from that of [3, 6]. We treat the nuisance parameters $K(f_0, x_0)$ and $r(\sigma)$ differently in the limit theory (1.3), with a particular view that *it is $K(f_0, x_0)$ that constitutes the main difficulty of using (1.3) for inference, while the problem of $r(\sigma)$ is relatively minor.* The underlying reason for this is that $K(f_0, x_0)$ involves the information for derivatives of f_0 at x_0, which cannot be obtained in a simple way from point estimates that take local averages, while $r(\sigma)$ can be relatively easily estimated using known methods (e.g., difference estimators [36, 54, 57]), or large samples (if available) in the data-driven local block $[\widehat{u}(x_0), \widehat{v}(x_0)]$.

The idea for the construction of the proposed CI (1.7) has a much broader scope of applications beyond the isotonic regression model (1.1). In Section 3, similar constructions of CIs are exploited in a number of other models with monotonicity shape constraints, including (i) monotone density estimation [26, 27, 56], (ii) interval censoring model with current status data [34], (iii) estimation of the mean function of a counting process with panel count data [66], and more generally, (iv) generalized linear models with monotonicity. These new CIs share the same general scheme that the constructions utilize the local information encoded by the analogue of $\widehat{u}(x_0), \widehat{v}(x_0)$ as in the regression setting, and require essentially no more efforts than performing the (maximum likelihood) estimation procedure once.

The results of this paper, in particular the pivotal limit distribution theory (1.8) and the resulting CI (1.7), make a significant further step in developing practical inference methods using the block estimators (1.2) beyond the limit theory (1.3) developed in [42], especially in view of the dependence of the constant factor $K(f_0, x_0)$ on the partial derivatives of the unknown f_0. However, the techniques used in proving (1.3) in [42] serve as the foundation for establishing the limit theory in (1.8) in this paper. In addition, as a key technical ingredient in proving (1.8), we show that the limiting Gaussian white noise versions of properly rescaled $\widehat{u}(x_0), \widehat{v}(x_0)$ are almost surely well defined (see Lemma A.1), so the limit distribution $\mathbb{L}_{\boldsymbol{\alpha}}$ in (1.8) is indeed well defined.

The rest of the article is organized as follows. In Section 2, we give a review of the limit distribution theory (1.3) developed in [42], study the proposed CI (1.7), and present the pivotal limit distribution theory (1.8). Some comparisons with the Banerjee–Wellner likelihood based inference method are also detailed in Section 2. In Section 3, we illustrate the generality of our method of constructing CIs in the four models mentioned above. Section 4 contains extensive simulation results that demonstrate the accuracy of the coverage probability of our proposed CIs, along with a detailed numerical comparison with the Banerjee–Wellner CIs [3, 6]. For clarity of presentation, proofs are deferred to the Appendix [14].

1.2. *Notation.* For the simplicity of presentation, we write the CI $[\widehat{\theta} - c_0, \widehat{\theta} + c_0]$ which is symmetric around $\widehat{\theta}$ as $\mathcal{I} = [\widehat{\theta} \pm c_0]$.

For a real-valued measurable function f defined on $(\mathcal{X}, \mathcal{A}, P)$, $\|f\|_{L_p(P)} \equiv \|f\|_{P,p} \equiv (P|f|^p)^{1/p}$ denotes the usual L_p-norm under P, and $\|f\|_\infty \equiv \sup_{x\in\mathcal{X}}|f(x)|$. Let $(\mathcal{F}, \|\cdot\|)$ be a subset of the normed space of real functions $f : \mathcal{X} \to \mathbb{R}$. For $\varepsilon > 0$ let $\mathcal{N}(\varepsilon, \mathcal{F}, \|\cdot\|)$ be the ε-covering number of $\mathcal{F}$; see [65], page 83, for more details.

For two real numbers a, b, $a \vee b \equiv \max\{a, b\}$ and $a \wedge b \equiv \min\{a, b\}$. For $x \in \mathbb{R}^d$, let $\|x\|_p$ denote its p-norm $(0 \le p \le \infty)$. For any $x, y \in \mathbb{R}^d$, $x \le y$ if and only if $x_i \le y_i$ for all $1 \le i \le d$. Let $[x, y] \equiv \prod_{k=1}^d [x_k \wedge y_k, x_k \vee y_k]$, $xy \equiv (x_k y_k)_{k=1}^d$, and $x \wedge (\vee) y \equiv (x_k \wedge (\vee) y_k)_{k=1}^d$. Let $\mathbf{1}_d = (1, \dots, 1) \in \mathbb{R}^d$. For $\ell_1, \ell_2 \in \{1, \dots, d\}$, we let $\mathbf{1}_{[\ell_1:\ell_2]} \in \mathbb{R}^d$ be such that $(\mathbf{1}_{[\ell_1:\ell_2]})_k = \mathbf{1}_{\ell_1 \le k \le \ell_2}$. We use C_x to denote a generic constant that depends only on x, whose numeric value may change from line to line unless otherwise specified. $a \lesssim_x b$ and $a \gtrsim_x b$ mean $a \le C_x b$ and $a \ge C_x b$ respectively, and $a \asymp_x b$ means $a \lesssim_x b$ and $a \gtrsim_x b$ [$a \lesssim b$ means $a \le Cb$ for some absolute constant C]. $\mathcal{O}_{\mathbf{P}}$ and $\mathfrak{o}_{\mathbf{P}}$ denote the usual big and small O notation in probability. $\rightsquigarrow$ is reserved for weak convergence. For two integers $k_1 > k_2$, we interpret $\sum_{k=k_1}^{k_2} \equiv 0$, $\prod_{k=k_1}^{k_2} \equiv 1$. We also interpret $(\infty)^{-1} \equiv 0$, $0/0 \equiv 0$.

2. Confidence interval: Isotonic regression.

2.1. *Limit distribution theory in* [42]: *A review.* Let us now describe the setting under which limit distribution theory for the block max–min and min–max estimators (1.2) is developed in [42]. The exposition below largely follows [42].

First, some further notation. For $f : \mathbb{R}^d \to \mathbb{R}$, and $k \in \{1, \dots, d\}$, $\alpha_k \in \mathbb{Z}_{\ge 1}$, let $\partial_k^{\alpha_k} f(x) \equiv \frac{d^{\alpha_k}}{dx_k^{\alpha_k}} f(x)$. For a multi-index $\boldsymbol{j} = (j_1, \dots, j_d) \in \mathbb{Z}_{\ge 0}^d$, let $\partial^{\boldsymbol{j}} \equiv \partial_1^{j_1} \cdots \partial_d^{j_d}$, and $\boldsymbol{j}! \equiv j_1! \cdots j_d!$ and $x^{\boldsymbol{j}} \equiv x_1^{j_1} \dots x_d^{j_d}$ for $x \in \mathbb{R}^d$. For $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_d) \in \mathbb{Z}_{\ge 1}^d$ in Assumption A below, that is, for some $0 \le s \le d$, $1 \le \alpha_1, \dots, \alpha_s < \infty = \alpha_{s+1} = \dots = \alpha_d$, let $J(\boldsymbol{\alpha})$ (resp. $J_*(\boldsymbol{\alpha})$) be the set of all $\boldsymbol{j} = (j_1, \dots, j_d) \in \mathbb{Z}_{\ge 0}^d$ satisfying $0 < \sum_{k=1}^s j_k/\alpha_k \le 1$ (resp. $\sum_{k=1}^s j_k/\alpha_k = 1$) and $j_k = 0$ for $s + 1 \le k \le d$, and let $J_0(\boldsymbol{\alpha}) \equiv J(\boldsymbol{\alpha}) \cup \{\mathbf{0}\}$. We often write $J = J(\boldsymbol{\alpha})$, $J_* = J_*(\boldsymbol{\alpha})$ and $J_0 = J_0(\boldsymbol{\alpha})$ if no confusion arises.

ASSUMPTION A. f_0 is coordinate-wise nondecreasing (i.e., $f_0 \in \mathcal{F}_d$), and is $\boldsymbol{\alpha}$-smooth at x_0 with intrinsic dimension s, $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_d)$ with integers $1 \le \alpha_1, \dots, \alpha_s < \infty = \alpha_{s+1} = \dots = \alpha_d$, $0 \le s \le d$, in the sense that $\partial_k^{j_k} f_0(x_0) = 0$ for $1 \le j_k \le \alpha_k - 1$ and $\partial_k^{\alpha_k} f_0(x_0) \ne 0$, $1 \le k \le s$, and in rectangles of the form $\bigcap_{k=1}^d \{|(x - x_0)_k| \le L_0 \cdot (r_n)_k\}$, $r_n = (\omega_n^{1/\alpha_1}, \dots, \omega_n^{1/\alpha_d})$ with $\omega_n > 0$, the Taylor expansion of f_0 satisfies for all $L_0 > 0$,

$$\lim_{\omega_n \downarrow 0} \omega_n^{-1} \sup_{\substack{x\in[0,1]^d,\\ |(x-x_0)_k| \le L_0\cdot(r_n)_k,\\ 1\le k\le d}} \left| f_0(x) - \sum_{\boldsymbol{j}\in J_0} \frac{\partial^{\boldsymbol{j}} f_0(x_0)}{\boldsymbol{j}!}(x - x_0)^{\boldsymbol{j}} \right| = 0.$$

The above assumption will be satisfied if f_0 depends only on its first s coordinates, and is locally $C^{\max_{1\le k\le s}\alpha_k}$ at x_0, with $\partial_k^{j_k} f_0(x_0) = 0$ for $1 \le j_k \le \alpha_k - 1$ and $\partial_k^{\alpha_k} f_0(x_0) \ne 0$, $1 \le k \le s$.

ASSUMPTION B. The design points $\{X_i\}_{i=1}^n$ satisfy either of the following:

- (*Fixed design*) $\{X_i\}$'s follow a $\boldsymbol{\beta}$-fixed lattice design: there exist some $\{\beta_1, \ldots, \beta_d\} \subset (0,1)$ with $\sum_{k=1}^d \beta_k = 1$ such that $x_0 \in \{X_i\}_{i=1}^n = \prod_{k=1}^d \{x_{1,k}, \ldots, x_{n_k,k}\}$, where $\{x_{1,k}, \ldots, x_{n_k,k}\}$ are equally spaced in $[0,1]$ (i.e., $|x_{j,k} - x_{j+1,k}| = 1/n_k$ for all $j = 1, \ldots, n_k - 1$) and $n_k = \lfloor n^{\beta_k} \rfloor$.
- (*Random design*) $\{X_i\}$'s follow i.i.d. uniform random design in $[0,1]^d$ and are also independent of $\{\xi_i\}$'s.

In $\boldsymbol{\beta}$-fixed lattice design, we assume without loss of generality

$$0 \le \alpha_1\beta_1 \le \cdots \le \alpha_s\beta_s \le \cdots \le \alpha_d\beta_d \le \infty. \tag{2.1}$$

Otherwise we may find a permutation of $\{1, \ldots, d\}$ to satisfy the above condition and the theory below will be carried over for the permuted indices.

In the random design, we assume for simplicity that the law P of X_i is uniform on $[0,1]^d$; the forthcoming Theorem 0 holds with minor changes when P is relaxed to have Lebesgue density π that is bounded away from 0 and ∞ on $[0,1]^d$ and is continuous over an open set containing the region $\{((x_0)_1, \ldots, (x_0)_s, x_{s+1}, \ldots, x_d) : 0 \le x_k \le 1, s+1 \le k \le d\}$. More discussion on the above assumptions is referred to [42]. The following limit distribution theory is obtained by [42].

THEOREM 0. *Let $x_0 \in (0,1)^d$. Suppose Assumptions A and B hold, and the errors $\{\xi_i\}$ are i.i.d. mean-zero with finite variance $\mathbb{E}\xi_1^2 = \sigma^2 < \infty$ (and are independent of $\{X_i\}$ in the random design case). Let κ_*, n_* be defined by*

	$\boldsymbol{\beta}$*-fixed lattice design*	*Random design*
κ_*	$\arg\max_{1\le\ell\le d} \frac{\sum_{k=\ell}^d \beta_k}{2+\sum_{k=\ell}^s \alpha_k^{-1}}$	1
n_*	$n^{\sum_{k=\kappa_*}^d \beta_k}$	n

If κ_* is uniquely defined, then for some finite random variables $\mathbb{C}^{\mp}(f_0, x_0)$,

$$(n_*/\sigma^2)^{\frac{1}{2+\sum_{k=\kappa_*}^s \alpha_k^{-1}}} \big(\widehat{f}_n^{\mp}(x_0) - f_0(x_0)\big) \rightsquigarrow \mathbb{C}^{\mp}(f_0, x_0).$$

Furthermore, if either $\{\alpha_k\}_{k=1}^s$ is a set of relative primes, that is, the greatest common divisor of $\{\alpha_{k_1}, \alpha_{k_2}\}$ is 1 for all $1 \le k_1 < k_2 \le s$, or all mixed derivatives $\partial^{\boldsymbol{j}} f_0$ of f_0 vanish at x_0 for all $\boldsymbol{j} \in J_*$, then

$$\mathbb{C}^{\mp}(f_0, x_0) =_d K(f_0, x_0) \cdot \mathbb{D}_{\boldsymbol{\alpha}}^{\mp},$$

where $K(f_0, x_0) = \{\prod_{k=\kappa_*}^s (\partial_k^{\alpha_k} f_0(x_0)/(\alpha_k+1)!)^{1/\alpha_k}\}^{\frac{1}{2+\sum_{k=\kappa_*}^s \alpha_k^{-1}}}$, and $\mathbb{D}_{\boldsymbol{\alpha}}^{\mp}$ are given by

$$\mathbb{D}_{\boldsymbol{\alpha}}^{-} \equiv \sup_{g_1 \in \mathscr{G}_1} \inf_{g_2 \in \mathscr{G}_2} \mathbb{V}_{\boldsymbol{\alpha}}(g_1, g_2), \qquad \mathbb{D}_{\boldsymbol{\alpha}}^{+} \equiv \inf_{g_2 \in \mathscr{G}_2} \sup_{g_1 \in \mathscr{G}_1} \mathbb{V}_{\boldsymbol{\alpha}}(g_1, g_2).$$

Here

$$\mathscr{G}_1 \equiv \{g_1 \in \mathbb{R}^d : g_1 > 0, (g_1)_k \le (x_0)_k, s+1 \le k \le d\},$$

$$\mathscr{G}_2 \equiv \{g_2 \in \mathbb{R}^d : g_2 > 0, (g_2)_k \le (1-x_0)_k, s+1 \le k \le d\},$$

$$\mathbb{V}_{\boldsymbol{\alpha}}(g_1, g_2) \equiv \frac{\mathbb{G}(g_1, g_2)}{\prod_{k=\kappa_*}^d ((g_1)_k + (g_2)_k)} + \sum_{k=\kappa_*}^s \frac{(g_2)_k^{\alpha_k+1} - (g_1)_k^{\alpha_k+1}}{(g_2)_k + (g_1)_k},$$

where $\mathbb{G}$ is a Gaussian process defined on $\mathbb{R}^d_{\geq 0} \times \mathbb{R}^d_{\geq 0}$ with the following covariance structure: for any (g_1, g_2), (g_1', g_2'),

$$\operatorname{Cov}\big(\mathbb{G}(g_1, g_2), \mathbb{G}(g_1', g_2')\big) = \prod_{k=\kappa_*}^{d} \big((g_1)_k \wedge (g_1')_k + (g_2)_k \wedge (g_2')_k\big).$$

Strictly speaking, [42] proved the case for the block max–min estimator $\widehat{f}_n^-$, but the case for the block min–max estimator $\widehat{f}_n^+$ follows from the same proofs. We refer the readers to [42] for detailed discussion and some concrete examples for κ_*, n_*. The merit of using the block average estimator $\widehat{f}_n$ is discussed in Section 4.3.

Theorem 0 is comprehensive under the general Assumptions A and B. To capture the essence, we consider a simple yet important case in the following Corollary 1.

COROLLARY 1. *Suppose f_0 is locally C^1 at $x_0 \in (0,1)^d$ with $\partial_k f_0(x_0) > 0$ for all $1 \leq k \leq d$, and the design points $\{X_i\}$ either (i) form a balanced fixed lattice design with $\beta_1 = \cdots = \beta_d = 1/d$, or (ii) follow a uniform random design on $[0,1]^d$. Suppose the errors $\{\xi_i\}$ are i.i.d. mean-zero with finite variance $\mathbb{E}\xi_1^2 = \sigma^2 < \infty$ (and are independent of $\{X_i\}$ in the random design case). Then,*

$$\big(n/\sigma^2\big)^{1/(2+d)}\big(\widehat{f}_n^{\mp}(x_0) - f_0(x_0)\big) \rightsquigarrow \left\{\prod_{k=1}^{d} \big(\partial_k f_0(x_0)/2\big)\right\}^{1/(2+d)} \cdot \mathbb{D}^{\mp}_{\mathbf{1}_d}.$$

REMARK 1. Theorem 0 (and Corollary 1) in the random design case requires the independence of the errors and the random design points due to the precise form of the limit distribution. Although this independence assumption is not most desirable, it is common in limit distribution theories for other one-dimensional shape-constrained regression estimators under random design settings, for instance in the convex regression model [24]. It is an interesting open question to see if the theory carries over to the more general settings, for instance $\mathbb{E}[\xi_i|X_i] = 0$ and $\mathbb{E}[\xi_i^2|X_i = x] = \sigma^2(x)$ for some smooth enough function σ.

2.2. *Pivotal limit distribution theory.* In this subsection, we formally establish the pivotal limit distribution theory (1.8). The main idea of the pivotal limit distribution theory (1.8) is that the information for $K(f_0, x_0)$ is already encoded in $\widehat{u}(x_0), \widehat{v}(x_0)$, so after proper scaling, we may naturally view $\sqrt{n_{\widehat{u},\widehat{v}}(x_0)/\sigma^2}$ as an estimator for $\{\omega_n(\boldsymbol{\alpha}) r(\sigma) K(f_0, x_0)\}^{-1} = (n_*/\sigma^2)^{1/(2+\sum_{k=\kappa_*}^{s} \alpha_k^{-1})}/K(f_0, x_0)$. Indeed, we have the following.

THEOREM 1. *Let $x_0 \in (0,1)^d$. Suppose Assumptions A and B hold, and the errors $\{\xi_i\}$ are i.i.d. mean-zero with finite variance $\mathbb{E}\xi_1^2 = \sigma^2 < \infty$ (and are independent of $\{X_i\}$ in the random design case). If the limit distribution in Theorem 0 is of the explicit form $\mathbb{C}^{\mp}(f_0, x_0) = K(f_0, x_0) \cdot \mathbb{D}^{\mp}_{\boldsymbol{\alpha}}$, then with $\widehat{f}_n(x_0)$ and $n_{\widehat{u},\widehat{v}}(x_0)$ defined respectively, in (1.5) and (1.6)*

$$\sqrt{n_{\widehat{u},\widehat{v}}(x_0)}\big(\widehat{f}_n(x_0) - f_0(x_0)\big) \rightsquigarrow \sigma \cdot \mathbb{L}_{\boldsymbol{\alpha}}.$$

Here $\mathbb{L}_{\boldsymbol{\alpha}}$ is a finite random variable defined by

$$\mathbb{L}_{\boldsymbol{\alpha}} \equiv \mathbb{S}_{\boldsymbol{\alpha}}\big(g^*_{1,\boldsymbol{\alpha}}, g^*_{2,\boldsymbol{\alpha}}\big) \cdot \frac{1}{2}\Big(\sup_{g_1 \in \mathscr{G}_1} \inf_{g_2 \in \mathscr{G}_2} \mathbb{V}_{\boldsymbol{\alpha}}(g_1, g_2) + \inf_{g_2 \in \mathscr{G}_2} \sup_{g_1 \in \mathscr{G}_1} \mathbb{V}_{\boldsymbol{\alpha}}(g_1, g_2)\Big),$$

where $g^*_{1,\boldsymbol{\alpha}}$ and $g^*_{2,\boldsymbol{\alpha}}$ are almost surely uniquely determined via

$$\inf_{g_2\in\mathscr{G}_2} \mathbb{V}_{\boldsymbol{\alpha}}(g^*_{1,\boldsymbol{\alpha}}, g_2) = \sup_{g_1\in\mathscr{G}_1}\inf_{g_2\in\mathscr{G}_2} \mathbb{V}_{\boldsymbol{\alpha}}(g_1, g_2),$$

$$\sup_{g_1\in\mathscr{G}_1} \mathbb{V}_{\boldsymbol{\alpha}}(g_1, g^*_{2,\boldsymbol{\alpha}}) = \inf_{g_2\in\mathscr{G}_2}\sup_{g_1\in\mathscr{G}_1} \mathbb{V}_{\boldsymbol{\alpha}}(g_1, g_2),$$

and

$$\mathbb{S}_{\boldsymbol{\alpha}}(g^*_{1,\boldsymbol{\alpha}}, g^*_{2,\boldsymbol{\alpha}}) \equiv \sqrt{\prod_{k=\kappa_*}^{s} (g^*_{2,\boldsymbol{\alpha}} + g^*_{1,\boldsymbol{\alpha}})_k},$$

with the $\mathscr{G}_i (i = 1, 2)$, $\mathbb{V}_{\boldsymbol{\alpha}}$ and κ_* in Theorem 0. In particular, $\mathbb{L}_{\boldsymbol{\alpha}}$ does not depend on $K(f_0, x_0)$.

Theorem 1 provides a *pivotal* limit distribution theory in the sense that the limit distribution of certain statistic concerning $\widehat{f}_n(x_0) - f_0(x_0)$ does not depend on the *difficult nuisance parameter $K(f_0, x_0)$*.

This pivotal phenomenon can be understood from an oracle perspective. As an illustration, we focus on the leading case $\boldsymbol{\alpha} = (1, \dots, 1)$ and $\sigma = 1$ in the setting of balanced fixed lattice design. In this case, $n_* = n$ and $\kappa_* = 1$. The oracle bandwidth vector $h^* = h^*(x_0)$ balances the bias and variance:

$$h^*_\ell \partial_\ell f_0(x_0) \approx \frac{1}{\sqrt{\prod_{k=1}^d (n^{1/d} h^*_k)}} \quad \text{for } 1 \le \ell \le d,$$

which yields

$$h^*_\ell \approx \big(\partial_\ell f_0(x_0)\big)^{-1} \cdot n^{-1/(2+d)} \left(\prod_{k=1}^d \partial_k f_0(x_0)\right)^{1/(2+d)}.$$

Hence, with $u^* = u^*(x_0) = x_0 - h^*/2$ and $v^* = v^*(x_0) = x_0 + h^*/2$,

$$\begin{aligned}
&\big|\sqrt{n_{u^*,v^*}(x_0)}\big(\widehat{f}_n(x_0) - f_0(x_0)\big)\big| \\
&\quad\approx \sqrt{n\prod_{k=1}^d (v^*_k - u^*_k) \cdot n^{-1/(d+2)} K(f_0, x_0) \cdot |\mathcal{O}_{\mathbf{P}}(1)|} \\
&\quad= \sqrt{n^{2/(2+d)}\left(\prod_{k=1}^d \partial_k f_0(x_0)\right)^{-2/(2+d)} \cdot n^{-1/(d+2)} K(f_0, x_0) \cdot |\mathcal{O}_{\mathbf{P}}(1)|} \\
&\quad= \text{const.} \times \big|\mathcal{O}_{\mathbf{P}}(1)\big|,
\end{aligned}$$

where $|\mathcal{O}_{\mathbf{P}}(1)|$ denotes a universal random variable. Theorem 1 can then be understood as the data driven bandwidth vectors $\widehat{u}(x_0), \widehat{v}(x_0)$ mimic the above oracle vectors $u^*(x_0), v^*(x_0)$ in achieving a pivotal limiting behavior. We note that the asymptotic variance of $\widehat{f}_n(x_0)$ is proportional to $(n^{-1}\prod_{k=1}^d \partial_k f_0(x_0))^{2/(2+d)}$.

2.3. *Confidence interval.* The pivotal limit distribution theory in Theorem 1 naturally implies the tuning free CI (1.7). In this subsection, we study (1.7) under fixed lattice and uniform random designs as in Assumption B. The nonuniform random design case will be discussed at the end.

To construct the CI, it remains to find a good estimate for the variance σ^2. Estimation of variance in the nonparametric regression model (1.1) is a well studied topic in the literature; for instance, we may use the class of difference estimators [36, 54, 57]. Below we present a ‘principled’ estimator that shows the reason why σ^2 is much easier to estimate than $K(f_0, x_0)$—point estimates already contain enough information for the variance, as long as a law of large numbers is satisfied. Formally, let

$$\sigma^2_{\widehat{u},\widehat{v}} \equiv \frac{1}{n_{\widehat{u},\widehat{v}}(x_0)} \sum_{X_i \in [\widehat{u}(x_0), \widehat{v}(x_0)]} \big(Y_i - \widehat{f}_n(x_0)\big)^2. \tag{2.2}$$

Note that $\sigma^2_{\widehat{u},\widehat{v}}$ only requires information of the observed $\{Y_i\}$ in the data driven neighborhood $[\widehat{u}(x_0), \widehat{v}(x_0)]$ of x_0 and the fitted value $\widehat{f}_n(x_0)$. Intuitively, since we have large samples in $[\widehat{u}(x_0), \widehat{v}(x_0)]$, it is natural to expect good performance of $\sigma^2_{\widehat{u},\widehat{v}}$ in the large sample limit. In fact, we have the following proposition.

PROPOSITION 1. *Under the conditions of Theorem 0, $\sigma^2_{\widehat{u},\widehat{v}} \to_p \sigma^2$.*

One theoretical advantage of (2.2) compared with the difference estimators is that $\sigma^2_{\widehat{u},\widehat{v}}$ takes local average around x_0, and therefore may *in principle* estimate the variance even in the heteroscedastic regression setting, for example, when the variance of $\bar{Y}|_{[u,v]}$ is given by $\sigma^2_{u,v}$ with a certain strictly positive and continuous $\sigma_{u,v}$ defined on the entire $[0, 1]^d \times [0, 1]^d$. The practical issue, however, is that the effective sample size in $[\widehat{u}(x_0), \widehat{v}(x_0)]$ is relatively small so (2.2) typically requires very large samples to achieve accurate variance estimation, in particular for $d \geq 2$.

With a consistent variance estimator, Theorem 1 can then be used to justify the use of the CI of the form defined in (1.7).

We first consider the leading case $\boldsymbol{\alpha} = \mathbf{1}_d$, where it is possible to construct *asymptotically exact CIs*.

THEOREM 2. *Let $c_\delta > 0$ be a continuity point of the d.f. of $|\mathbb{L}_{\mathbf{1}_d}|$ such that*

$$\mathbb{P}\big(|\mathbb{L}_{\mathbf{1}_d}| > c_\delta\big) = \delta. \tag{2.3}$$

For any consistent variance estimator $\widehat{\sigma}^2$, that is, $\widehat{\sigma}^2 \to_p \sigma^2$, the CI $\mathcal{I}_n(c_\delta)$ defined in (1.7) satisfies

$$\lim_{n\to\infty} \mathbb{P}_{f_0}\big(f_0(x_0) \in \mathcal{I}_n(c_\delta)\big) = 1 - \delta. \tag{2.4}$$

Furthermore, with $K(f_0, x_0) = (\prod_{k=1}^d (\partial_k f_0(x_0)/2))^{1/(d+2)}$ as in Theorem 0, for any $\varepsilon > 0$,

$$\liminf_{n\to\infty} \mathbb{P}_{f_0}\big(|\mathcal{I}_n(c_\delta)| < 2c_\delta \mathfrak{g}_\varepsilon \cdot (\sigma^2/n)^{1/(d+2)} K(f_0, x_0)\big) \geq 1 - \varepsilon. \tag{2.5}$$

Here $\mathfrak{g}_\varepsilon \in (0, \infty)$ is such that

$$\mathbb{P}\big(\mathbb{S}^{-1}_{\mathbf{1}_d} \geq \mathfrak{g}_\varepsilon\big) \leq \varepsilon. \tag{2.6}$$

Theorem 2 shows that if the critical value c_δ is chosen according to (2.3), then the CI $\mathcal{I}_n(c_\delta)$ achieves the asymptotic exact confidence level $1 - \delta$. Furthermore, the CI $\mathcal{I}_n(c_\delta)$ shrinks at the oracle length, being automatically adaptive to the unknown information on the derivatives of f_0 at x_0, that is, $K(f_0, x_0)$.

The choice of the critical value c_δ depends on the distribution of $\mathbb{L}_{\mathbf{1}_d}$. Since $\mathbb{L}_{\mathbf{1}_d}$ does not depend on the unknown regression function f_0, it is possible to simulate c_δ for different values of confidence levels $\delta > 0$. See Section 4 for more details on simulated critical values of $\mathbb{L}_{\mathbf{1}_d}$ for $d = 1, 2, 3$.

Note that the CI in Theorem 2, although enjoying the advantage of achieving asymptotically exact confidence level, is not adaptive to the unknown local smoothness $\boldsymbol{\alpha}$ of the isotonic regression function f_0 at x_0. By relaxing the requirement of *exact* CI and calibrating the critical value c_δ only, the following theCI $\mathcal{I}_n(c_\delta)$ may adapt to all local smoothness level $\boldsymbol{\alpha}$ as well as the unknown information of f_0 at x_0 expressed by $K(f_0, x_0)$. More concretely, we have the following theorem.

THEOREM 3. *Let $c_\delta > 0$ be chosen such that*

$$\sup_{\boldsymbol{\alpha}} \mathbb{P}(|\mathbb{L}_{\boldsymbol{\alpha}}| > c_\delta) \le \delta. \tag{2.7}$$

For any consistent variance estimator $\widehat{\sigma}^2$, that is, $\widehat{\sigma}^2 \to_p \sigma^2$, the CI $\mathcal{I}_n(c_\delta)$ defined in (1.7) satisfies

$$\liminf_{n\to\infty} \mathbb{P}_{f_0}\big(f_0(x_0) \in \mathcal{I}_n(c_\delta)\big) \ge 1 - \delta. \tag{2.8}$$

Furthermore, with κ_*, n_* and $K(f_0, x_0)$ defined in Theorem 0, for any $\varepsilon > 0$,

$$\liminf_{n\to\infty} \mathbb{P}_{f_0}\big(\big|\mathcal{I}_n(c_\delta)\big| < 2c_\delta \mathfrak{g}_{\varepsilon,\boldsymbol{\alpha}}\big(\sigma^2/n_*\big)^{\frac{1}{2+\sum_{k=\kappa_*}^{s}\alpha_k^{-1}}} K(f_0, x_0)\big) \ge 1 - \varepsilon. \tag{2.9}$$

Here $\mathfrak{g}_{\varepsilon,\boldsymbol{\alpha}} \in (0, \infty)$ is such that

$$\mathbb{P}\big(\mathbb{S}_{\boldsymbol{\alpha}}^{-1}(g_{1,\boldsymbol{\alpha}}^*, g_{2,\boldsymbol{\alpha}}^*) \ge \mathfrak{g}_{\varepsilon,\boldsymbol{\alpha}}\big) \le \varepsilon. \tag{2.10}$$

To make the above theorem useful for construction of $\boldsymbol{\alpha}$-adaptive CIs, it is crucial to choose the critical value $c_\delta > 0$ such that (2.7) is satisfied. The proposition below shows that this is indeed possible.

PROPOSITION 2. *The following holds for some constant $L_0 > 0$ depending only on d, x_0: for any $t \ge 1$,*

$$\sup_{\boldsymbol{\alpha}} \mathbb{P}\big(|\mathbb{L}_{\boldsymbol{\alpha}}| > t\big) \le L_0 \exp\big(-t^{4/(d+2)}/L_0\big).$$

Hence, it suffices to choose $c_\delta \asymp_{d,x_0} \log^{(d+2)/4}(1/\delta)$ to satisfy (2.7).

Nonuniform random design. So far we have assumed that the design distribution P is uniform on $[0, 1]^d$ in the random design case. The situation will be more complicated for general design distributions P, as the limit distribution in Theorem 0 can depend on P in a rather complicated and nonlocal way. Note that for general P, Theorem 0 requires the Lebesgue density π of P to be bounded away from 0 and ∞ on $[0, 1]^d$ and to be continuous in the neighborhood of x_0. In the special case where $s = d$ (i.e., $f_0(x)$ depends on all elements of x at $x = x_0$), the effect of P is local and can be factored out in the limit distribution theory in Theorem 0; see [42], Remark 1 (5), for detailed discussion. In this case, using similar arguments as in the proof of Theorem 1, we still have

$$\sqrt{n_{\widehat{u},\widehat{v}}(x_0)}\big(\widehat{f}_n(x_0) - f_0(x_0)\big) \rightsquigarrow \sigma \cdot \mathbb{L}_{\mathbf{1}_d}.$$

Hence, for any consistent variance estimator $\widehat{\sigma}^2$, (1.7) continues to be an asymptotic $1 - \delta$ CI of $f_0(x_0)$ which shrinks at the oracle length in a similar sense to the statements of Theorem 2.

2.4. *Comparison with the approach of* [6] *in* $d = 1$. [5, 6] developed an inference procedure using the log likelihood ratio test for various monotone-response models in the univariate case $d = 1$. In the regression setting, this idea is best illustrated in the random design, with P being the uniform distribution on $[0, 1]$ and the errors $\{\xi_i\}$ being i.i.d. $\mathcal{N}(0, 1)$. The block max–min and min–max estimators $\widehat{f}_n^{\mp}$ defined via (1.2) and their average $\widehat{f}_n$ all reduce to the univariate LSE at design points.

We consider testing the hypothesis

$$H_0 : f_0(x_0) = m_0 \quad \text{vs.} \quad H_1 : f_0(x_0) \neq m_0.$$

To form a likelihood ratio test, let $\widehat{f}_n^0$ be the *constrained* least squares estimator defined via

$$\widehat{f}_n^0 \in \operatorname*{argmin}_{f \in \mathcal{F}_1, f(x_0)=m_0} \sum_{i=1}^n \big(Y_i - f(X_i)\big)^2.$$

$\widehat{f}_n^0$ is well defined on the design points, and can be computed by performing two isotonic regressions on $\{(Y_i, f(X_i)) : X_i \leq x_0\}$ and $\{(Y_i, f(X_i)) : X_i > x_0\}$ followed by thresholding (see [3], page 939). Under the Gaussian likelihood, the likelihood ratio test statistic is given by

$$2 \log \lambda_n(m_0) = -\sum_{i=1}^n \big(Y_i - \widehat{f}_n(X_i)\big)^2 + \sum_{i=1}^n \big(Y_i - \widehat{f}_n^0(X_i)\big)^2. \tag{2.11}$$

[5, 6] showed that if f_0 is locally C^1 at x_0 with $f_0'(x_0) > 0$ and H_0 holds, then

$$2 \log \lambda_n(m_0) \rightsquigarrow \mathbb{K}_1, \tag{2.12}$$

where the distribution $\mathbb{K}_1$ is free of the nuisance parameter $f_0'(x_0)$ that would otherwise be present in the limit distribution theory (cf. Theorem 0 in the simplest case $d = 1, \alpha = 1$). A CI of $f_0(x_0)$ can now be obtained through inversion of (2.12): Let $\mathcal{I}_n^{\mathrm{BW}}(d_\delta) \equiv \{m_0 : 2 \log \lambda_n(m_0) \leq d_\delta\}$ (BW refers to Banerjee–Wellner) with $\mathbb{P}(\mathbb{K}_1 > d_\delta) = \delta$. Then

$$\mathbb{P}_{f_0}\big(f_0(x_0) \in \mathcal{I}_n^{\mathrm{BW}}(d_\delta)\big) \to \mathbb{P}(\mathbb{K}_1 \leq d_\delta) = 1 - \delta.$$

REMARK 2. For a fixed m_0, the likelihood ratio test requires two isotonic regressions to calculate the test statistic (2.11), which can be computed efficiently thanks to the fast PAVA algorithms. With carefully written algorithms the inversion of (2.12) may not add too much computational burden to obtain the likelihood ratio test based CIs; see, for example, [28] for a fast algorithm in the related current status model. However, the proposed procedure (1.7) is still computationally simpler and more straightforward.

The validity of the Banerjee–Wellner CI crucially relies on the assumption $f_0'(x_0) > 0$. Compared with our procedure, it is natural to wonder if the likelihood ratio approach offers a stronger degree of universality in terms of adaptation to unknown local smoothness of the regression function. As we will show below, the limit distribution of the log likelihood ratio test statistic does depend on the number of vanishing derivatives of f_0, and is therefore not adaptive to the local smoothness of f_0.

To formally state the result, let $\mathrm{slogcm}(f, I)$ be the left-hand slope of the greatest convex minorant of f restricted to the interval I. Write $\mathrm{slogcm}(f) = \mathrm{slogcm}(f, \mathbb{R})$ for simplicity. Let

$$\mathrm{slogcm}^0(f) = \big(\mathrm{slogcm}(f, (-\infty, 0]) \wedge 0\big)\mathbf{1}_{(-\infty,0]} + \big(\mathrm{slogcm}(f, (0, \infty)) \vee 0\big)\mathbf{1}_{(0,\infty)}.$$

Let $\mathbb{B}$ be the standard two-sided Brownian motion started from 0, and $X_{a,b;\alpha}(t) \equiv a\mathbb{B}(t) + bt^{\alpha+1}$. Let $g_{a,b;\alpha} \equiv \mathrm{slogcm}(X_{a,b;\alpha})$ and $g^0_{a,b;\alpha} \equiv \mathrm{slogcm}^0(X_{a,b;\alpha})$. These quantities are a.s.

well defined for $b > 0$ and an odd integer $\alpha \geq 1$ as $X_{a,b;\alpha}(t)$ is of the order $\mathcal{O}_{a.s.}(t^{\alpha+1})$ for $t \to \pm\infty$ and is a.s. bounded on compacta. Informally, $g_{a,b;\alpha}$ is the 'isotonic regression for $X_{a,b;\alpha}$' in the Gaussian white noise model $\mathrm{d}X_{a,b;\alpha}(t) = b(\alpha+1)t^\alpha\,\mathrm{d}t + a\,\mathrm{d}\mathbb{B}(t)$, and $g^0_{a,b;\alpha}$ is the 'constrained isotonic regression' subject to $g^0_{a,b;\alpha}(0) = 0$.

THEOREM 4. *Consider the above setting. Suppose f_0 is nondecreasing and locally C^α at x_0 for some $\alpha \geq 1$, with $\partial^j f_0(x_0) = 0$ for $j = 1, \ldots, \alpha - 1$ and $\partial^\alpha f_0(x_0) \neq 0$. Then under H_0,*

$$2\log\lambda_n(m_0) \rightsquigarrow \int_{\mathbb{R}} \{(g_{1,1;\alpha}(t))^2 - (g^0_{1,1;\alpha}(t))^2\}\,\mathrm{d}t \equiv \mathbb{K}_\alpha.$$

REMARK 3. By [42], Lemma 1, α's satisfying the assumption of Theorem 4 must be odd, and $\partial^\alpha f_0(x_0) > 0$.

It is clear from Theorem 4 that the limit distribution for the log likelihood ratio test depends on the unknown local smoothness level α of f_0 through the slope processes $g_{1,1;\alpha}, g^0_{1,1;\alpha}$. This phenomenon is observed numerically in [18] in another related setting: the limit distributions for the log-likelihood ratio tests for the mode of a log-concave density depend on the number of vanishing derivatives at the mode.

REMARK 4. If the variance σ^2 is unknown, then the log-likelihood ratio test statistic involves the unknown σ^2. By taking (2.11) as the definition of the quantity of $2\log\lambda_n(m_0)$, it holds (under the same conditions as in Theorem 4) that $2\log\lambda_n(m_0) \rightsquigarrow \sigma^2 \cdot \mathbb{K}_\alpha$.

3. Beyond isotonic regression: Inference in other monotone models.

The idea for constructing tuning-free CIs in the previous section has a much broader scope beyond the setting of multiple isotonic regression. As a proof of concept, in this section we construct CIs for a few further nonparametric models with certain monotonicity shape constraints, adapting essentially the same idea as developed in the previous section.

3.1. *The common scheme.* We briefly outline the common scheme for the construction of CIs in the models to be studied in detail below. Suppose we want to estimate a univariate monotone function f_0. There is a natural piecewise constant estimator $\widehat{f}_n$ (usually the maximum likelihood estimator) for f_0 that exhibits a nonstandard limit distribution at the point of interest x_0, typically at a cube-root rate $\omega_n = n^{-1/3}$ under curvature conditions on f_0' at x_0:

$$\begin{aligned}\omega_n^{-1}\big(\widehat{f}_n(x_0) - f_0(x_0)\big) &\rightsquigarrow \sup_{h_1>0}\inf_{h_2>0}\bigg[a \cdot \frac{\mathbb{G}(h_1,h_2)}{h_1+h_2} + b\cdot(h_2 - h_1)\bigg] \\ &=_d \inf_{h_2>0}\sup_{h_1>0}\bigg[a \cdot \frac{\mathbb{G}(h_1,h_2)}{h_1+h_2} + b\cdot(h_2 - h_1)\bigg] \\ &=_d (a^2 b)^{1/3}\cdot\mathbb{D}_1.\end{aligned}$$

Recall that $\mathbb{D}_1 \equiv \mathbb{D}_1^+ =_d \mathbb{D}_1^-$ is defined in Theorem 0. In fact, $\mathbb{D}_1/2$ follows the well-known Chernoff distribution (cf. [30]).

Here we have two nuisance parameters, namely a, b:

- b is a *difficult* nuisance parameter to estimate, which usually involves the derivatives of the monotone function to be estimated. This will be tackled by the analogue of '$n_{\widehat{u},\widehat{v}}(x_0)$'.
- a is typically *easy* to estimate, either via observed samples or via fitted values in the analogue of the 'local block $[\widehat{u}(x_0), \widehat{v}(x_0)]$' of x_0.

One special feature in the one-dimensional setting is the exchangability of supremum and infimum in the limit distribution, so $\widehat{u}(x_0)$ and $\widehat{v}(x_0)$ are simply the left and right end-points of the constant piece of $\widehat{f}_n$ that contains x_0. As $\widehat{v}(x_0)-\widehat{u}(x_0)$ is of order $\mathcal{O}_{\mathbf{P}}(\omega_n(a/b)^{2/3})$, we would expect the following pivotal limit distribution theory:

$$\sqrt{n\big(\widehat{v}(x_0)-\widehat{u}(x_0)\big)}\big(\widehat{f}_n(x_0)-f_0(x_0)\big)\rightsquigarrow a\cdot\mathbb{L}_1. \tag{3.1}$$

Now given a consistent estimate $\widehat{a}_n$ of a, we have the following generic CI of $f_0(x_0)$:

$$\mathcal{I}_n^*(c_\delta)\equiv\big[\widehat{f}_n(x_0)\pm c_\delta\cdot\widehat{a}_n/\sqrt{n\big(\widehat{v}(x_0)-\widehat{u}(x_0)\big)}\big].$$

Similar to the regression setting, the construction of CIs in specific models to be detailed below is tuning-free, and requires essentially no further efforts beyond a single step of (maximum likelihood) estimation.

3.2. *Monotone density estimation.* Consider the classical problem of estimating a decreasing density f_0 on $[0,\infty)$ based on i.i.d. observations $X_1,\dots,X_n$. The maximum likelihood estimator (MLE) $\widehat{f}_n$, known as the *Grenander estimator*, is the left derivative of the least concave majorant of the empirical distribution function $\mathbb{F}_n$. By the max–min representation, for any $x_0\in(0,\infty)$, we may write

$$\widehat{f}_n(x_0)=\inf_{0<u<x_0}\sup_{v\geq x_0}\frac{\mathbb{F}_n(v)-\mathbb{F}_n(u)}{v-u}=\frac{\mathbb{F}_n(\widehat{v}(x_0))-\mathbb{F}_n(\widehat{u}(x_0))}{\widehat{v}(x_0)-\widehat{u}(x_0)},$$

where $(\widehat{u}(x_0),\widehat{v}(x_0))$ is the a.s. uniquely specified pair for which the last equality in the above display holds. It is well known (see, e.g., [26, 27, 55, 65]) that if f_0 is locally C^1 at x_0 with $f_0'(x_0)<0$ and $f_0(x_0)>0$, then

$$\begin{aligned}n^{1/3}\big(\widehat{f}_n(x_0)-f_0(x_0)\big)&\rightsquigarrow\sup_{h_1>0}\inf_{h_2>0}\bigg[\sqrt{f_0(x_0)}\cdot\frac{\mathbb{G}(h_1,h_2)}{h_1+h_2}+\frac{1}{2}\big|f_0'(x_0)\big|(h_2-h_1)\bigg]\\&=_d\big(f_0(x_0)\big|f_0'(x_0)\big|/2\big)^{1/3}\cdot\mathbb{D}_1.\end{aligned}$$

To use the above limit theorem to form CI, the difficult nuisance to estimate is $f_0'(x_0)$, while the easy one is $f_0(x_0)$. The inference problem in the density setting is recently tackled in [31], using both the log likelihood ratio test approach similar to [3, 6] and a bootstrap assisted approach for the smoothed maximum likelihood estimator. Our proposal for a CI of $f_0(x_0)$ is the following:

$$\mathcal{I}_n^{\mathrm{den}}(c_\delta)\equiv\big[\widehat{f}_n(x_0)\pm c_\delta\cdot\sqrt{\widehat{f}_n(x_0)}/\sqrt{n\big(\widehat{v}(x_0)-\widehat{u}(x_0)\big)}\big]\cap[0,\infty).$$

Let $\mathbb{L}_1$ and $\mathbb{S}_1$ be as in Theorem 1 with $d=1$ and $\boldsymbol{\alpha}=1$.

THEOREM 5. *Suppose f_0 is locally C^1 at x_0 with $f_0'(x_0)<0$ and $f_0(x_0)>0$. Let $c_\delta>0$ be a continuity point of the d.f. of $|\mathbb{L}_1|$ such that $\mathbb{P}(|\mathbb{L}_1|>c_\delta)=\delta$. Then*

$$\lim_{n\to\infty}\mathbb{P}_{f_0}\big(f_0(x_0)\in\mathcal{I}_n^{\mathrm{den}}(c_\delta)\big)=1-\delta.$$

Furthermore, for any $\varepsilon>0$,

$$\liminf_{n\to\infty}\mathbb{P}_{f_0}\big(\big|\mathcal{I}_n^{\mathrm{den}}(c_\delta)\big|<2c_\delta\mathfrak{g}_\varepsilon\cdot n^{-1/3}\big(f_0(x_0)\big|f_0'(x_0)\big|/2\big)^{1/3}\big)\geq1-\varepsilon.$$

Here $\mathfrak{g}_\varepsilon\in(0,\infty)$ is such that $\mathbb{P}(\mathbb{S}_1^{-1}\geq\mathfrak{g}_\varepsilon)\leq\varepsilon$.

It is also possible to consider adaptive CIs by calibrating the critical value c_δ similarly as in Theorem 3. We omit the details.

3.3. *Current status data*: *Interval censoring model.* Let $X_1, \ldots, X_n$ and $T_1, \ldots, T_n$ be independent i.i.d. samples from distribution functions F_0 and G_0 supported on $[0, \infty)$. Let $\Delta_i \equiv \mathbf{1}_{X_i \le T_i}$. We observe $(\Delta_1, T_1), \ldots, (\Delta_n, T_n)$ and want to estimate F_0, the distribution of unobserved $X_1, \ldots, X_n$. Consider the maximum likelihood estimator $\widehat{F}_n$ that maximizes

$$F \mapsto \sum_{i=1}^n \bigl(\Delta_i \log F(T_i) + (1 - \Delta_i) \log\bigl(1 - F(T_i)\bigr)\bigr). \tag{3.2}$$

Let $T_{(1)} \le \cdots \le T_{(n)}$ be the order statistics of $T_1, \ldots, T_n$. It is well known (see, e.g., [34, 65]) that the solutions $(\widehat{F}_n(T_{(1)}), \ldots, \widehat{F}_n(T_{(n)}))$ is given by the isotonic regression over $(\Delta_{(i)} = \mathbf{1}_{X_{(i)} \le T_{(i)}})_{i=1}^n$. In other words, for any $t_0 \in (0, \infty)$,

$$\widehat{F}_n(t_0) = \max_{i: T_i \le t_0} \min_{j: T_j \ge t_0} \frac{\sum_{k=i}^j \Delta_{(k)}}{j - i + 1} \equiv \max_{u \le t_0} \min_{v \ge t_0} \bar{\Delta}_{(\cdot)}|_{[u,v]} = \bar{\Delta}_{(\cdot)}|_{[\widehat{u}(t_0), \widehat{v}(t_0)]},$$

where $(\widehat{u}(t_0), \widehat{v}(t_0))$ is any pair for which the last equality in the above display holds. It is also well known (see, e.g., [34, 65]) that if F_0, G_0 has positive and locally continuous density f_0, g_0 at t_0, then

$$\begin{aligned} & n^{1/3}\bigl(\widehat{F}_n(t_0) - F_0(t_0)\bigr) \\ & \quad \rightsquigarrow \sup_{h_1 > 0} \inf_{h_2 > 0} \biggl[\sqrt{F_0(t_0)\bigl(1 - F_0(t_0)\bigr)/g_0(t_0)} \cdot \frac{\mathbb{G}(h_1, h_2)}{h_1 + h_2} + \frac{1}{2} f_0(t_0)(h_2 - h_1)\biggr] \\ & \quad =_d \bigl(F_0(t_0)\bigl(1 - F_0(t_0)\bigr) f_0(t_0)/2g_0(t_0)\bigr)^{1/3} \cdot \mathbb{D}_1. \end{aligned}$$

The inference problem in the current status model is investigated in [6, 31] using likelihood ratio methods; see [4] for similar likelihood ratio based inference methods in the context of monotone, uni-modal and U–shaped failure rates under a right–censoring mechanism. Here we take a different approach, similar to our proposal in the regression setting. Note that the difficult nuisance parameter in this problem is $f_0(t_0)$ since $X_1, \ldots, X_n$ are unobserved, while $F_0(t_0)$ and $g_0(t_0)$ are easy to estimate. For instance, we may use $\widehat{F}_n(t_0)$ to estimate $F_0(t_0)$, and

$$\widehat{g}_n(t_0) \equiv \sum_i \mathbf{1}_{T_i \in [\widehat{u}(t_0), \widehat{v}(t_0)]} / \bigl\{n\bigl(\widehat{v}(t_0) - \widehat{u}(t_0)\bigr)\bigr\}$$

to estimate $g_0(t_0)$. Now consider the following CI for $F_0(t_0)$:

$$\begin{aligned} \mathcal{I}_n^{\mathrm{cur}}(c_\delta) & \equiv \bigl[\widehat{F}_n(t_0) \pm c_\delta \cdot \sqrt{\widehat{F}_n(t_0)\bigl(1 - \widehat{F}_n(t_0)\bigr)/\widehat{g}_n(t_0)} / \sqrt{n\bigl(\widehat{v}(t_0) - \widehat{u}(t_0)\bigr)}\bigr] \cap [0, 1] \\ & = \Biggl[\widehat{F}_n(t_0) \pm c_\delta \cdot \sqrt{\widehat{F}_n(t_0)\bigl(1 - \widehat{F}_n(t_0)\bigr)} / \sqrt{\sum_i \mathbf{1}_{T_i \in [\widehat{u}(t_0), \widehat{v}(t_0)]}}\Biggr] \cap [0, 1]. \end{aligned}$$

THEOREM 6. *Suppose F_0, G_0 has positive and locally continuous density f_0, g_0 at t_0. Let $c_\delta > 0$ be a continuity point of the d.f. of $|\mathbb{L}_1|$ such that $\mathbb{P}(|\mathbb{L}_1| > c_\delta) = \delta$. Then*

$$\lim_{n \to \infty} \mathbb{P}_{F_0, G_0}\bigl(F_0(t_0) \in \mathcal{I}_n^{\mathrm{cur}}(c_\delta)\bigr) = 1 - \delta.$$

Furthermore, for any $\varepsilon > 0$,

$$\begin{aligned} & \liminf_{n \to \infty} \mathbb{P}_{F_0, G_0}\bigl(\bigl|\mathcal{I}_n^{\mathrm{cur}}(c_\delta)\bigr| \\ & \qquad < 2 c_\delta \mathfrak{g}_\varepsilon \cdot n^{-1/3}\bigl(F_0(t_0)\bigl(1 - F_0(t_0)\bigr) f_0(t_0)/2g_0(t_0)\bigr)^{1/3}\bigr) \ge 1 - \varepsilon. \end{aligned}$$

Here $\mathfrak{g}_\varepsilon \in (0, \infty)$ is such that $\mathbb{P}(\mathbb{S}_1^{-1} \ge \mathfrak{g}_\varepsilon) \le \varepsilon$.

3.4. *Panel count data*: *Counting process model.* The examples in previous subsections are amongst the 'classical' ones in the field of monotonicity-constrained estimation. Below we consider one further example, in the context of *panel count data*, that is less 'classical' due to its increased complexity. The inference problem for this model is previous studied in [61] using likelihood ratio methods.

Here is the setup. We follow the notation in [66]. Suppose that $N = \{N(t) : t \geq 0\}$ is a counting process with mean function $\Lambda_0(t) = \mathbb{E}N(t)$. Let K be an integer-valued random variable, and $T = \{T_{k,j} : 1 \leq j \leq k, k \geq 1\}$ be an triangular array of observation times. We assume that N and (K, T) are independent and $T_{k,j-1} \leq T_{k,j}$. Let $X = (N_K, T_K, K)$, where $T_K = (T_{K,1}, \ldots, T_{K,K})$ and $N_K = (N(T_{K,1}), \ldots, N(T_{K,K}))$. We observe i.i.d. copies $X_1, \ldots, X_n$ of X, where $X_i = (N^{(i)}_{K_i}, T^{(i)}_{K_i}, K_i)$. The problem is to estimate $\Lambda_0(t)$. By building a Poisson model for $N(t) \sim_d \text{Poisson}(\Lambda_0(t))$, and pretending independence of the events within each sample X_i, we may consider the estimator $\widehat{\Lambda}_n$ that maximizes the pseudo log-likelihood

$$\Lambda \mapsto \sum_{i=1}^{n}\sum_{j=1}^{K_i}\big[N^{(i)}_{K_i,j} \log \Lambda\big(T^{(i)}_{K_i,j}\big) - \Lambda\big(T^{(i)}_{K_i,j}\big)\big]. \tag{3.3}$$

Let $s_1 < s_2 < \cdots < s_m$ be the ordered distinct observation time points in the set $\{T^{(i)}_{K_i,j} : 1 \leq j \leq K_i, i = 1, \ldots, n\}$. For $1 \leq \ell \leq m$, define

$$w_\ell \equiv \sum_{i=1}^{n}\sum_{j=1}^{K_i} \mathbf{1}_{T^{(i)}_{K_i,j}=s_\ell}, \qquad \bar{N}_\ell \equiv \frac{1}{w_\ell}\sum_{i=1}^{n}\sum_{j=1}^{K_i} N^{(i)}_{K_i,j}\mathbf{1}_{T^{(i)}_{K_i,j}=s_\ell}.$$

It is known (see, e.g., [64, 66]) that

$$\widehat{\Lambda}_n(t_0) = \max_{s_i \leq t_0}\min_{s_j \geq t_0} \frac{\sum_{p=i}^{j} w_p \bar{N}_p}{\sum_{p=i}^{j} w_p} = \frac{\sum_{p:\widehat{u}(t_0)\leq s_p \leq \widehat{v}(t_0)} w_p \bar{N}_p}{\sum_{p:\widehat{u}(t_0)\leq s_p \leq \widehat{v}(t_0)} w_p},$$

where $(\widehat{u}(t_0), \widehat{v}(t_0))$ is any pair in $\{s_1, \ldots, s_m\}^2$ such that the right-hand side of the above display holds. Under the assumption that Λ_0 is nondecreasing and locally C^1 with $\Lambda_0'(t_0) > 0$ and further regularity conditions, [66] proved the following limit distribution theory for $\widehat{\Lambda}_n(t_0)$:

$$\begin{aligned} & n^{1/3}\big(\widehat{\Lambda}_n(t_0) - \Lambda_0(t_0)\big) \\ & \quad \rightsquigarrow \sup_{h_1>0}\inf_{h_2>0}\bigg[\sqrt{\sigma^2(t_0)/g(t_0)} \cdot \frac{\mathbb{G}(h_1, h_2)}{h_1 + h_2} + \frac{1}{2}\Lambda_0'(t_0)(h_2 - h_1)\bigg] \\ & \quad =_d \big(\sigma^2(t_0)\Lambda_0'(t_0)/2g(t_0)\big)^{1/3} \cdot \mathbb{D}_1. \end{aligned}$$

Here $\sigma^2(t_0) \equiv \text{Var}(N(t_0))$ and $g(t_0) \equiv \sum_{k=1}^{\infty} \mathbb{P}(K = k)\sum_{j=1}^{k} g_{k,j}(t_0)$ with $g_{k,j}$ denoting the Lebesgue density of $T_{k,j}$.

The difficult nuisance parameter in this problem is $\Lambda_0'(t_0)$, and easier ones are $\sigma^2(t_0)$ and $g(t_0)$. For instance, with $n_{\widehat{u},\widehat{v}}(t_0)$ being the number of $\{T^{(i)}_{K_i,j}\}$ in the interval $[\widehat{u}(t_0), \widehat{v}(t_0)]$, that is,

$$n_{\widehat{u},\widehat{v}}(t_0) = \sum_{i=1}^{n}\sum_{j=1}^{K_i} \mathbf{1}_{T^{(i)}_{K_i,j} \in [\widehat{u}(t_0), \widehat{v}(t_0)]},$$

let $\widehat{g}_n(t_0) \equiv n_{\widehat{u},\widehat{v}}(t_0)/\{n(\widehat{v}(t_0) - \widehat{u}(t_0))\}$ and

$$\widehat{\sigma}_n^2(t_0) \equiv \frac{1}{n_{\widehat{u},\widehat{v}}(t_0)}\sum_{i=1}^{n}\sum_{j=1}^{K_i}\big(N^{(i)}_{K_i,j} - \widehat{\Lambda}_n(t_0)\big)^2 \mathbf{1}_{T^{(i)}_{K_i,j} \in [\widehat{u}(t_0), \widehat{v}(t_0)]}.$$

Consider the following CI for $\Lambda_0(t_0)$:

$$\begin{aligned}\mathcal{I}_n^{\mathrm{pan}}(c_\delta) &\equiv \Big[\widehat{\Lambda}_n(t_0) \pm c_\delta \cdot \sqrt{\widehat{\sigma}_n^2(t_0)/\widehat{g}_n(t_0)}\Big/\sqrt{n\big(\widehat{v}(t_0) - \widehat{u}(t_0)\big)}\Big] \cap [0, \infty) \\ &= \big[\widehat{\Lambda}_n(t_0) \pm c_\delta \cdot \widehat{\sigma}_n(t_0)/\sqrt{n_{\widehat{u},\widehat{v}}(t_0)}\big] \cap [0, \infty).\end{aligned}$$

THEOREM 7. *Suppose Λ_0 is nondecreasing and locally C^1 with $\Lambda_0'(t_0) > 0$ and further regularity conditions* (*as specified in Theorem* 4.3 *of* [66]) *hold. Let $c_\delta > 0$ be a continuity point of the d.f. of $|\mathbb{L}_1|$ such that $\mathbb{P}(|\mathbb{L}_1| > c_\delta) = \delta$. Then*

$$\lim_{n\to\infty} \mathbb{P}_{\Lambda_0}\big(\Lambda_0(t_0) \in \mathcal{I}_n^{\mathrm{pan}}(c_\delta)\big) = 1 - \delta.$$

Furthermore, for any $\varepsilon > 0$,

$$\liminf_{n\to\infty} \mathbb{P}_{\Lambda_0}\big(\big|\mathcal{I}_n^{\mathrm{pan}}(c_\delta)\big| < 2c_\delta \mathfrak{g}_\varepsilon \cdot n^{-1/3}\big(\sigma^2(t_0)\Lambda_0'(t_0)/2g(t_0)\big)^{1/3}\big) \geq 1 - \varepsilon.$$

Here $\mathfrak{g}_\varepsilon \in (0, \infty)$ is such that $\mathbb{P}(\mathbb{S}_1^{-1} \geq \mathfrak{g}_\varepsilon) \leq \varepsilon$.

REMARK 5. Similar to [66], we do not assume that the Poisson model for the counting process N, used in building the pseudo likelihood for the definition $\widehat{\Lambda}_n$, need to be true.

3.5. *Generalized linear models and the i.n.i.d.* (*independent, not identically distributed*) *case.* The likelihoods in (3.2) and (3.3) hint that a similar idea could be taken further to the generalized linear models as follows. Suppose real-valued random variables Y_i's ($i = 1, \ldots, n$) are independent samples with density $f(\cdot; \theta_{0,i})$ (with respect to a σ-finite measure ν on the real line) from a canonical exponential family:

$$f(y; \theta) = \exp\big(y \cdot p(\theta) - q(\theta)\big), \quad \theta \in \Theta, \tag{3.4}$$

where $\Theta = \{\theta \in \mathbb{R} : \int e^{y\cdot p(\theta)}\nu(\mathrm{d}y) < \infty\}$ and ν does not put all the mass at a single point y_0. We assume that $\theta_{0,i} \equiv \theta_0(x_i)$, where $\theta_0 : [0, 1] \to \mathbb{R}$ is monotonically nondecreasing. Let Θ_0 be the interior of Θ. In general we may assume that the natural parameter $p(\theta)$ is a continuously differentiable and strictly increasing function of θ. However, as the mean function $\mu(\theta) = \int yf(y; \theta)\nu(\mathrm{d}y) = \partial q(\theta)/\partial p(\theta)$ is always continuously differentiable and strictly increasing in $p(\theta)$ in Θ_0, we consider for simplicity the parametrization $\theta = \mu(\theta)$, as alternative parametrizations can be easily handled by applying the delta-method to our results. In this setting, the variance is given by $\int(y - \theta)^2 f(y; \theta)\nu(\mathrm{d}y) = 1/p'(\theta)$ in Θ_0. We shall assume that the variance is finite at $\theta = \theta_0(0)$ and $\theta = \theta_0(1)$ even when they are on the boundary of the domain Θ, for example, $p'(\theta_0(0)) = p'(\theta_0(1)) = \infty$ when $Y_i \in \{0, 1\}$.

We are interested in estimating θ_0 by the maximum likelihood estimator $\widehat{\theta}_n$ that maximizes

$$\theta \mapsto \sum_{i=1}^{n}\big(Y_i \cdot p\big(\theta(x_i)\big) - q\big(\theta(x_i)\big)\big)$$

over $\theta \in \mathbb{R}^n$ such that $\theta_1 \leq \cdots \leq \theta_n$. As $p(\theta)$ and $q(\theta)$ are analytic in Θ_0, by [58], Theorem 1.5.2, the solution $\widehat{\theta}_n \in \mathbb{R}^n$ is given by the isotonic regression of $(Y_i)_{i=1}^n$. For any $x \in [0, 1]$, let

$$\widehat{\theta}_n(x) \equiv \max_{i:x_i\leq x}\ \min_{j:x_j\geq x} \frac{\sum_{k=i}^{j} Y_k}{j - i + 1} \equiv \max_{u\leq x}\ \min_{v\geq x} \bar{Y}|_{[u,v]}. \tag{3.5}$$

Then we may identify $\widehat{\theta}_{n,i} = \widehat{\theta}_n(x_i)$, $i = 1, \ldots, n$.

From here the analysis of the maximum likelihood isotonic regression in the generalized linear model reduces to a special case of the analysis of the max–min estimator in the i.n.i.d. case where

$$Y_i \text{ are independent with } \mathbb{E}[Y_i] = \theta_0(x_i) \text{ and } \mathrm{Var}(Y_i) = \sigma^2(x_i) \tag{3.6}$$

under a Lindeberg condition on Y_i and a smoothness condition on $\theta_0(x_i)$.

Formally, let $x_1 \le \cdots \le x_n$ with $x_{i_0-1} \le x_0 \le x_{i_0}$ for some $i_0 \in \{2, \ldots, n-1\}$, $\alpha > 0$ be fixed and $\omega_n \equiv n^{-\alpha/(2\alpha+1)}$. For $(h_1, h_2) \in \mathbb{R}^2_{\ge 0}$ define $S_{n,h_1,h_2} \equiv \{i : i_0 - n\omega_n^{1/\alpha} h_1 \le i \le i_0 + n\omega_n^{1/\alpha} h_2\}$. We assume that for some $g_0(x_0) > 0$

$$\begin{gathered}\left| \sum_{i \in S_{n,h_1,h_2}} \frac{\theta_0(x_i) - \theta_0(x_0)}{\omega_n |S_{n,h_1,h_2}|} - g_0(x_0) \frac{h_2^{\alpha+1} - h_1^{\alpha+1}}{h_1 + h_2} \right| = \mathfrak{o}(1), \\ \left| \sum_{i \in S_{n,h_1,h_2}} \frac{\sigma^2(x_i)/\sigma^2(x_0)}{|S_{n,h_1,h_2}|} - 1 \right| = \mathfrak{o}(1), \\ \sum_{i \in S_{n,h_1,h_2}} \frac{\mathbb{E}[(Y_i - \theta_0(x_i))^2 \mathbf{1}_{(Y_i - \theta_0(x_i))^2 > c^{-1}\sigma^2(x_0)|S_{n,h_1,h_2}|}]}{\sigma^2(x_0)|S_{n,h_1,h_2}|} = \mathfrak{o}(1),\end{gathered} \tag{3.7}$$

uniformly in $(h_1, h_2) \in [1/c, c]^2$ for every $c > 1$. Under these conditions, we have the following theorem.

THEOREM 8.

1. *Suppose* (3.6) *and* (3.7) *hold with a nondecreasing function* θ_0 *and* $\max_{1 \le i \le n} \sigma^2(x_i) = \mathcal{O}(1)$. *Let* $\widehat{\theta}_n$ *be as in* (3.5). *Then*

$$\begin{aligned}&\omega_n^{-1}(\widehat{\theta}_n(x_0) - \theta_0(x_0)) \\ &\rightsquigarrow \sup_{h_1>0} \inf_{h_2>0} \left[\sigma(x_0) \cdot \frac{\mathbb{G}(h_1, h_2)}{h_1 + h_2} + g_0(x_0) \frac{h_2^{\alpha+1} - h_1^{\alpha+1}}{h_1 + h_2} \right] \\ &=_d \left((\sigma(x_0))^{2\alpha} g_0(x_0) \right)^{1/(2\alpha+1)} \cdot \mathbb{D}_\alpha.\end{aligned}$$

2. *Suppose* $\alpha\beta$ *is a positive odd integer for some* $\beta > 0$ *and that* $\theta_0(\cdot)$ *has* $\alpha\beta - 1$ *vanishing derivatives and positive the* $(\alpha\beta)$*-th derivative at* x_0. *Let* $\pi(\cdot)$ *be a density such that* $\pi(x) = (1 + \mathfrak{o}(1))\pi_0 \beta \cdot |x - x_0|^{\beta-1}$ *uniformly in a neighborhood of* x_0. *Then, the first line of* (3.7) *holds* (*in probability*) *when* $x_1 \le \cdots \le x_n$ *are the ordered independent samples from* $\pi(\cdot)$. *Moreover, in the generalized linear model* (3.4), *the second and third lines of* (3.7) *and the uniform boundedness condition on the variance always hold.*

In addition to providing a general limit distribution theory in the i.n.i.d. case, the above theorem specifies sufficient conditions under which the fast convergence rate with $\alpha > 1$ can be achieved when more x_i are sampled near x_0 than the usual $x_i = i/n$, and vice versa.

In Theorem 8, the difficult nuisance parameter is $g_0(x_0)$, and the easier one is $\sigma^2(x_0)$. Let $(\widehat{u}(x_0), \widehat{v}(x_0))$ be any pair such that $\widehat{\theta}_n(x_0) = \bar{Y}|_{[\widehat{u}(x_0), \widehat{v}(x_0)]}$. Consider the following CI for $\theta_0(x_0)$:

$$\mathcal{I}_n^{\mathrm{GLM}}(c_\delta) \equiv \left[\widehat{\theta}_n(x_0) \pm c_\delta \cdot \widehat{\sigma}_n \varrho_n / \sqrt{n(\widehat{v}(x_0) - \widehat{u}(x_0))}\right],$$

where $\widehat{\sigma}_n^2 \equiv 1/p'(\widehat{\theta}_n(x_0))$ under (3.4) or $\widehat{\sigma}_n^2 \equiv \sigma^2_{\widehat{u},\widehat{v}}$ as in (2.2) in general, and $\varrho_n = 1 + \mathfrak{o}_{\mathbf{P}}(1)$. If we choose $\varrho_n \equiv \sqrt{n(\widehat{v}(x_0) - \widehat{u}(x_0)) / \sum_i \mathbf{1}_{x_i \in [\widehat{u}(x_0), \widehat{v}(x_0)]}}$, the CI above reduces to $[\widehat{\theta}_n(x_0) \pm c_\delta \cdot \widehat{\sigma}_n / \sqrt{\sum_i \mathbf{1}_{x_i \in [\widehat{u}(x_0), \widehat{v}(x_0)]}}]$ in the similar form to (1.7).

TABLE 1
Simulated critical values c_δ for $d = 1$

δ	0.01	0.02	0.05	0.1	0.15	0.2	0.5
$f_0(x) = 2(x - 0.5)$	3.04	2.65	2.11	1.68	1.42	1.23	0.59
$f_0(x) = 5(x - 0.5)$	3.04	2.65	2.11	1.68	1.42	1.23	0.59
$f_0(x) = 10x^2$	3.04	2.66	2.11	1.68	1.42	1.23	0.59

THEOREM 9. *Assume the same conditions as in Theorem 8 with $\alpha = 1$. Let $c_\delta > 0$ be a continuity point of the d.f. of $|\mathbb{L}_1|$ such that $\mathbb{P}(|\mathbb{L}_1| > c_\delta) = \delta$. Then*

$$\lim_{n\to\infty} \mathbb{P}_{\theta_0}\big(\theta_0(x_0) \in \mathcal{I}_n^{\mathrm{GLM}}(c_\delta)\big) = 1 - \delta.$$

Furthermore, for any $\varepsilon > 0$,

$$\liminf_{n\to\infty} \mathbb{P}_{\theta_0}\big(\big|\mathcal{I}_n^{\mathrm{GLM}}(c_\delta)\big| < 2c_\delta \mathfrak{g}_\varepsilon \cdot n^{-1/3}\big(\sigma^2(x_0)g_0(x_0)\big)^{1/3}\big) \geq 1 - \varepsilon.$$

Here $\mathfrak{g}_\varepsilon \in (0, \infty)$ is such that $\mathbb{P}(\mathbb{S}_1^{-1} \geq \mathfrak{g}_\varepsilon) \leq \varepsilon$.

The above theorem covers only the usual case of $\alpha = 1$ for the cube-root rate. The critical value c_δ for general α can be handled as in Theorem 3.

4. Simulation studies.

4.1. *Critical values c_δ via simulations.* In this subsection, we discuss (i) simulation methods to approximate critical values c_δ of the pivotal limit distribution theory (with local smoothness $\boldsymbol{\alpha} = (1, \ldots, 1)$ as in Theorem 2), and (ii) data-driven adjustments to edge effect and small sample size.

4.1.1. *Simulated critical values c_δ.* We use the following method to simulate critical values c_δ:

1. Specify a true mean function f_0 defined on a fixed lattice $\{X_i\}$ in $[0, 1]^d$, and generate $B = 10^6$ repeated observations $\{\{Y_{i,b} = f_0(X_i) + \xi_{i,b}, i = 1, \ldots, n\} : b = 1, \ldots, B\}$ with i.i.d. $\xi_{i,b} \sim \mathcal{N}(0, \sigma^2)$ (we take $\sigma = 1$ for this simulation).
2. At design point x_0, we obtain $T(x_0; b) \equiv \{\sqrt{n_{\widehat{u},\widehat{v}}(x_0; b)}|\widehat{f}_n(x_0; b) - f_0(x_0)|/\sigma : b = 1, \ldots, B\}$, where $\widehat{f}_n(x_0; b)$ and $n_{\widehat{u},\widehat{v}}(x_0; b)$ are calculated via (1.5) and (1.6). We note that the block $[\widehat{u}(x_0), \widehat{v}(x_0)]$ is specified by $\widehat{u}(x_0)$ from the block max–min estimator and $\widehat{v}(x_0)$ from the block min–max estimator.
3. Find critical values c_δ by the corresponding quantiles of $\{T(x_0; b) : b = 1, \ldots, B\}$.

($d = 1$). The simulated critical values c_δ for $d = 1$ are summarized in Table 1. As the block max–min and min–max estimators (1.2) are equivalent to the isotonic least squares estimator (LSE) at design points in $d = 1$, we use `isoreg` (based on the PAVA algorithm [7, 58]) built in `R`. Let $n = 10^5$, so that $X_i = i/n$ for all $1 \leq i \leq n$ and $x_0 = 0.5$.

As the sample size ($n = 10^5$) for $d = 1$ is quite large, the estimates for different f_0's remain the same at least up to two decimal places, with an exception for $c_{0.02}$. Such precision is typically sufficient for the purpose of inference.

($d \geq 2$). For multiple isotonic regression, we use brute force to compute the block max–min and min–max estimators, which seems to be the only algorithm readily available. With computational complexity $\mathcal{O}(n^3)$, the brute force algorithm is much more expensive than the

TABLE 2
Simulated critical values c_δ for $d = 2$

δ	0.01	0.02	0.05	0.10	0.15	0.20	0.50
$f_0(x) = 2x_1 + 2x_2 - 2$	2.61	2.26	1.78	1.41	1.19	1.03	0.49
$f_0(x) = 2x_1 + 5x_2 - 3.5$	2.63	2.27	1.80	1.43	1.21	1.04	0.50
$f_0(x) = 5x_1 + 5x_2 - 5$	2.64	2.29	1.81	1.43	1.21	1.05	0.50

linear-time PAVA algorithm specific to the univariate isotonic LSE. Thus, it is not computationally feasible to perform $B = 10^6$ simulations on, say, a $10^5 \times 10^5$ (i.e., $n = 10^{10}$) lattice.

Nevertheless, we present below in Table 2 ($d = 2$) and Table 3 ($d = 3$) some encouraging simulation results by brute force computation over relatively small lattices:

- For $d = 2$, we use a 50×50 lattice and take sample critical values at $x_0 = (0.5, 0.5)$.
- For $d = 3$, we use a $16 \times 16 \times 16$ lattice and take sample critical values at $x_0 = (0.5, 0.5, 0.5)$.

The simulated critical values c_δ in $d = 2, 3$, albeit of small sample size in each dimension, already support the pivotal limit distribution theory. Their concrete numeric values are, however, less stable compared with $d = 1$ for different mean functions f_0, largely due to the curse of dimensionality that requires much larger sample size n to achieve similar accuracy as in $d = 1$. Unfortunately, brute force seems not ideal for this task. It is therefore of great interest to develop fast algorithms for the block max–min and min–max estimators (1.2) in view of their theoretically attractive properties.

By taking average, we give a few suggested critical values as follows.

4.1.2. *Data-driven adjustments.* As the pivotal limit distribution theory relies on local smoothness of f_0 and a large sample size, it is not surprising that for outskirt design points or when the sample size is relatively small, CIs constructed via (1.7) with the critical values suggested in Table 4 would be less accurate. See, for instance, the plots given below in Section 4.2 for demonstration. This is particularly relevant for $d \geq 2$, since a lot more points are present on the outskirts and in practice the sample size in each dimension is usually not as large as in the univariate case. These issues call for critical value adjustments to improve accuracy in inference.

Our proposal is to adjust the critical values based on the observed $\{Y_i\}$ and the sample size. More specifically, we propose the use of critical values simulated through a smooth proxy $\widehat{f}_{\mathrm{smooth}}$ of the block average estimator $\widehat{f}_n$. In order to match the noise level of $\{Y_i\}$, the variance σ^2 in simulation can be chosen to be the variance estimate $\widehat{\sigma}^2$ of $\{Y_i\}$. A simple smoothing method to get $\widehat{f}_{\mathrm{smooth}}$ is the isotonization of the LOESS fit $\widehat{f}_{\mathrm{loess}}$ (with default smoothing parameter built in R) of $\widehat{f}_n$, that is, $\widehat{f}_{\mathrm{smooth}} =$ the block average estimate for $\{\widehat{f}_{\mathrm{loess}}(X_i)\}$. See [53] for more details on constrained smoothing. As $\widehat{f}_{\mathrm{smooth}}$ is expected to be

TABLE 3
Simulated critical values c_δ for $d = 3$

δ	0.01	0.02	0.05	0.10	0.15	0.20	0.50
$f_0(x) = 2x_1 + 2x_2 + 2x_3 - 3$	2.26	1.96	1.55	1.24	1.05	0.91	0.44
$f_0(x) = 2x_1 + 5x_2 + 5x_3 - 6$	2.41	2.09	1.66	1.33	1.13	0.98	0.48
$f_0(x) = 5x_1 + 5x_2 + 5x_3 - 7.5$	2.41	2.10	1.67	1.34	1.14	0.99	0.49

TABLE 4
Suggested critical values c_δ (: use with caution)*

δ	$d = 1$	$d = 2$	$d = 3$
0.05	2.11	1.80*	1.63*
0.10	1.68	1.42*	1.30*

'close' to the true mean function of interest, it is reasonable to expect the simulated critical values for $\widehat{f}_{\mathrm{smooth}}$ to better mimic those for f_0 for design points on the outskirts and when the sample size is relatively small. We call the critical values simulated from $\widehat{f}_{\mathrm{smooth}}$ the *adjusted critical values*.

As will be clear from the next subsection, the *adjusted critical values* improve the inference accuracy both for design points on the outskirts and for smaller samples.

4.2. *Numerical performance of the proposed confidence intervals.* In this subsection, we investigate the numerical performance of the proposed CIs, exclusively in the multiple isotonic regression model. More specifically, we construct CIs for $f_0(x)$ at each design point $x \in \{X_i\}$ and compute their corresponding coverage probabilities as follows:

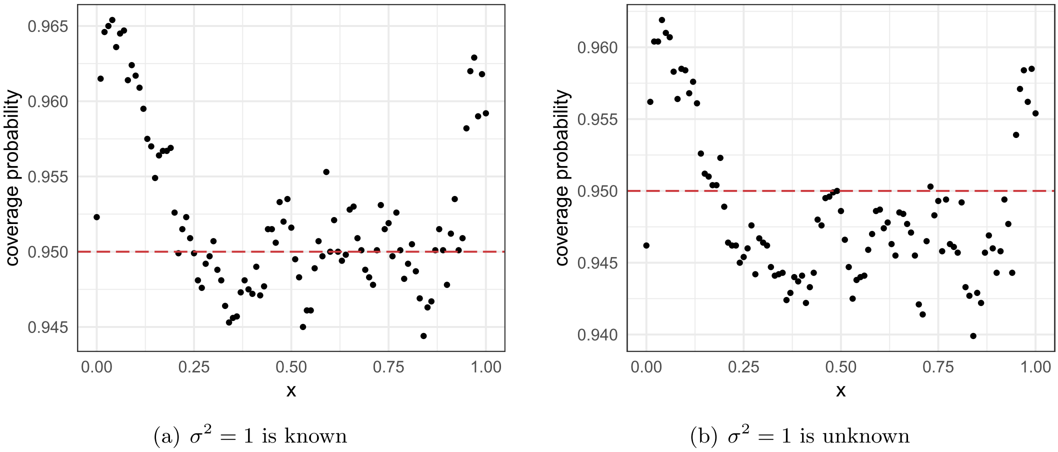
1. For each specified mean function f_0, generate $B = 10^4$ repeated observations $\{\{Y_{i,b} = f_0(X_i) + \xi_{i,b}, i = 1, \ldots, n\} : b = 1, \ldots, B\}$ with i.i.d. $\xi_{i,b} \sim \mathcal{N}(0, \sigma^2)$.
2. For each $b = 1, \ldots, B$, construct the CI $\mathcal{I}_n(x; c_\delta, b)$ for each design point x via (1.7). The CIs with the suggested c_δ in Table 4 are referred to as *vanilla CIs*, and the CIs with adjusted c_δ (as described in the proceeding subsection) as *CV-adjusted CIs*.
3. Report $B^{-1}\sum_{b=1}^{B} \mathbf{1}(f_0(x) \in \mathcal{I}_n(x; c_\delta, b))$ as the coverage probability at design point x, that is, the proportion of the CIs $\{\mathcal{I}_n(x; c_\delta, b) : b = 1, \ldots, B\}$ that successfully cover the truth $f_0(x)$ out of $B = 10^4$ repeated observations. We focus on 95% CIs, that is, $\delta = 0.05$.

For variance estimation, we use the class of difference estimators [36, 54, 57] rather than the principled estimator in (2.2), as the latter requires large samples that are computationally expensive for $d \geq 2$. Specifically, we use the following variance estimator $\widehat{\sigma}^2$:

$$\widehat{\sigma}^2 = \begin{cases} \sum_i (2Y_i - Y_{i-1} - Y_{i+1})^2 / (6(n-2)), & d = 1, \\ \sum_{i,j} (4Y_{i,j} - Y_{i-1,j} - Y_{i+1,j} - Y_{i,j-1} - Y_{i,j+1})^2 & \\ \quad / (20 \times (n_1 - 2)(n_2 - 2)), & d = 2, \\ \sum_{i,j,k} (6Y_{i,j,k} - Y_{i-1,j,k} - Y_{i+1,j,k} & \\ \quad - Y_{i,j-1,k} - Y_{i,j+1,k} - Y_{i,j,k-1} - Y_{i,j,k+1})^2 & \\ \quad / (42(n_1 - 2)(n_2 - 2)(n_3 - 2)), & d = 3, \end{cases} \tag{4.1}$$

where, with slight abuse of notation, the observations are $(Y_i)_{1 \leq i \leq n}$ for $d = 1$, $(Y_{i,j})_{1 \leq i \leq n_1, 1 \leq j \leq n_2}$ for $d = 2$, and $(Y_{i,j,k})_{1 \leq i \leq n_1, 1 \leq j \leq n_2, 1 \leq k \leq n_3}$ for $d = 3$.

4.2.1. *Coverage probability.* The scatter plots of coverage probabilities at all design points in $d = 1$ are given in Figure 2. In $d = 1$, we consider $f_0(x) = e^{2x}$ and $n = 100$, so that $x \in \{0.01, 0.02, \ldots, 1.00\}$. We observe slightly larger errors of the coverage probabilities of the vanilla CIs at points near $x = 0$ or $x = 1$, but overall the coverage errors are small for the small sample size $n = 100$. When σ^2 is unknown and estimated by the difference estimator (4.1), the errors are slightly inflated at most of the design points, but are still controlled



(a) $\sigma^2 = 1$ is known

(b) $\sigma^2 = 1$ is unknown

FIG. 2. *Scatter plots for the coverage probabilities of the* 95% *CIs in* $d = 1$, *where* $f_0(x) = e^{2x}$ *and* $n = 100$.

within 1%. The CIs are overall biased slightly towards under-coverage, which is possibly due to the bias in variance estimation: the median of $\widehat{\sigma}^2$'s from 10^4 simulations is 0.9803, slightly smaller than the true $\sigma^2 = 1$.

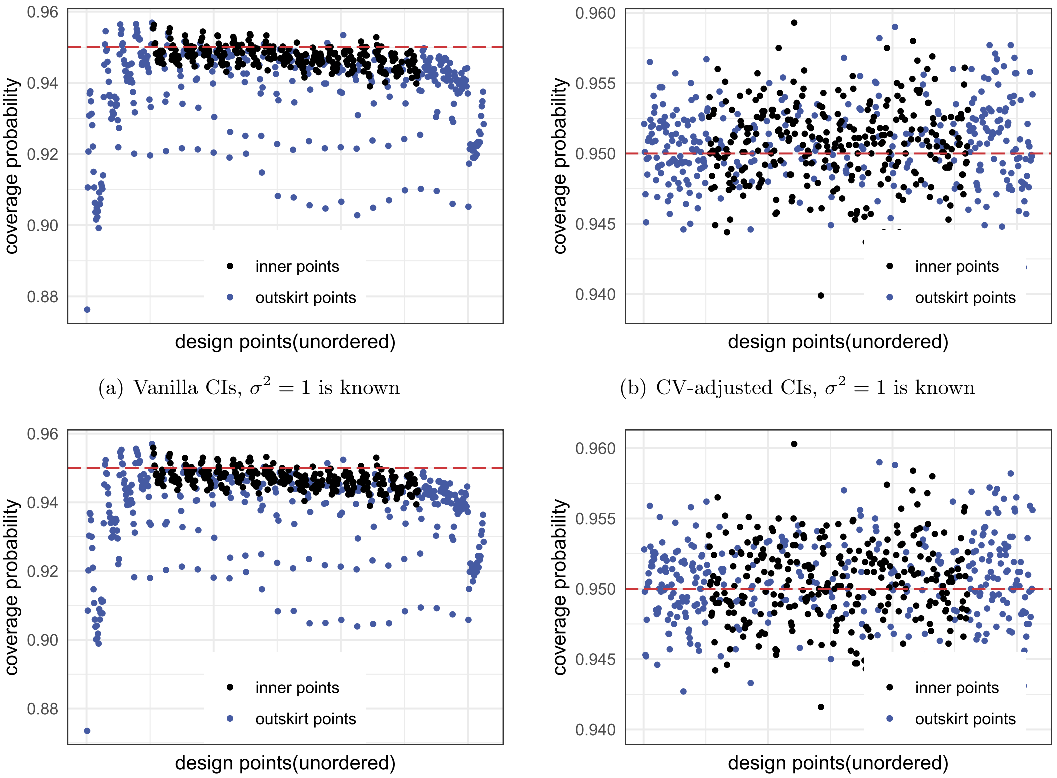
The scatter plots of coverage probabilities in $d = 2$ are given in Figure 3. We consider test function $f_0(x) = e^{x_1+x_2}$ on a 25×25 lattice on $[0, 1]^2$ so that $(x_1, x_2) \in \{(i/25, j/25) : i = 1, \dots, 25, j = 1, \dots, 25\}$. We call design points in the inner 17×17 lattice (i.e., $\{x : 5/25 \le x_1 \le 21/25, 5/25 \le x_2 \le 21/25\}$) inner points, and the rest outskirt points. We can clearly identify edge effect in Figure 3 when using the approximated universal critical value 1.80 in Table 4; the coverage probabilities at outskirt points are more biased as shown in Figure 3(a) and (c). The CV-adjusted CIs significantly reduce the edge effect and improve the coverage accuracy, as shown in Figure 3(b) and (d). Figure 4 shows the boxplots for the coverage probabilities at all design points using the aforementioned two types of CIs (vanilla and CV-adjusted) and under both known and unknown σ^2. The CV-adjusted CIs clearly have more accurate coverage.

Our simulation results for $d = 3$ exhibit similar phenomena to the case $d = 2$. See the scatter plots in Figure 5 and the boxplots in Figure 6, where we designate the design points in the inner $5 \times 5 \times 5$ lattice as inner points and the rest outskirt points. For the vanilla CIs, we use the approximated universal critical value 1.63 in Table 4. The lattice of size $9 \times 9 \times 9$ has too few points in each dimension, so distributional approximation to the pivot limit is less accurate, as shown in Figure 5(a) and (c). On the other hand, the CV-adjusted CIs yield much better empirical results, as shown in Figure 5(b) and (d).

In conclusion, the above simulations support our proposed inference procedure via the pivotal limit distribution theory. In situations when sample size (in terms of each dimension) is relatively small, or the design points are on the outskirts, CV-adjusted CIs are shown to significantly improve the inference accuracy compared with the vanilla CIs that use universal critical values in Table 4.

4.2.2. *CI lengths*. Another important theoretical property stated in Theorem 2 is that the proposed CI (1.7) shrinks at the oracle rate. Suppose we know the partial derivatives $\{\partial_k f_0(x_0), 1 \le k \le d\}$, the limit distribution theory in Corollary 1 implies an oracle CI

$$\left[\widehat{f}_n(x_0) \pm c_\delta \cdot \left(n/\sigma^2\right)^{-1/(d+2)} \left\{\prod_{k=1}^{d} \left(\partial_k f_0(x_0)/2\right)\right\}^{1/(2+d)}\right], \tag{4.2}$$



(a) Vanilla CIs, $\sigma^2 = 1$ is known

(b) CV-adjusted CIs, $\sigma^2 = 1$ is known

(c) Vanilla CIs, $\sigma^2 = 1$ is unknown

(d) CV-adjusted CIs, $\sigma^2 = 1$ is unknown

FIG. 3. *Scatter plots for the coverage probabilities of the* 95% *CIs in* $d = 2$, *where* $f_0(x) = e^{x_1+x_2}$ *in* 25×25 *lattice.*

where c_δ is the $1 - \delta$ quantile of $|\mathbb{D}| = |(\mathbb{D}^-_{\mathbf{1}_d} + \mathbb{D}^+_{\mathbf{1}_d})/2|$ which can be similarly simulated as in Section 4.1.1. Recall $\mathbb{D}^{\mp}_1/2$ follows Chernoff distribution, so $c_{0.05} = 1.9964$ in $d = 1$; see, for example, [30], Table 3.1. Our simulation suggests that $c_{0.05}$ is approximately 1.85 in $d = 2$ and 1.78 in $d = 3$. Then, Theorem 2 (2.5) asserts that the length of the proposed CI should shrink at the same rate as the length of the oracle CI in (4.2).

To see this in finite samples, we carry out a simulation that follows the same procedure as before but with varying sample sizes n. Only balanced fixed lattice design is considered in this simulation, so sample size n indicates an $n^{1/d} \times \cdots \times n^{1/d}$ lattice. See Figure 7 for the boxplots for the lengths of the proposed CI based on 10^4 simulations, where the lengths

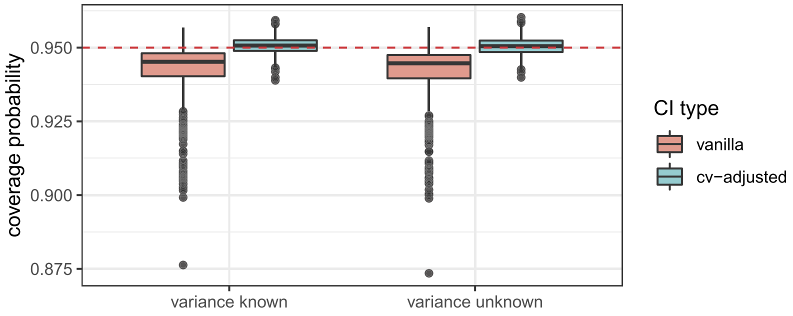
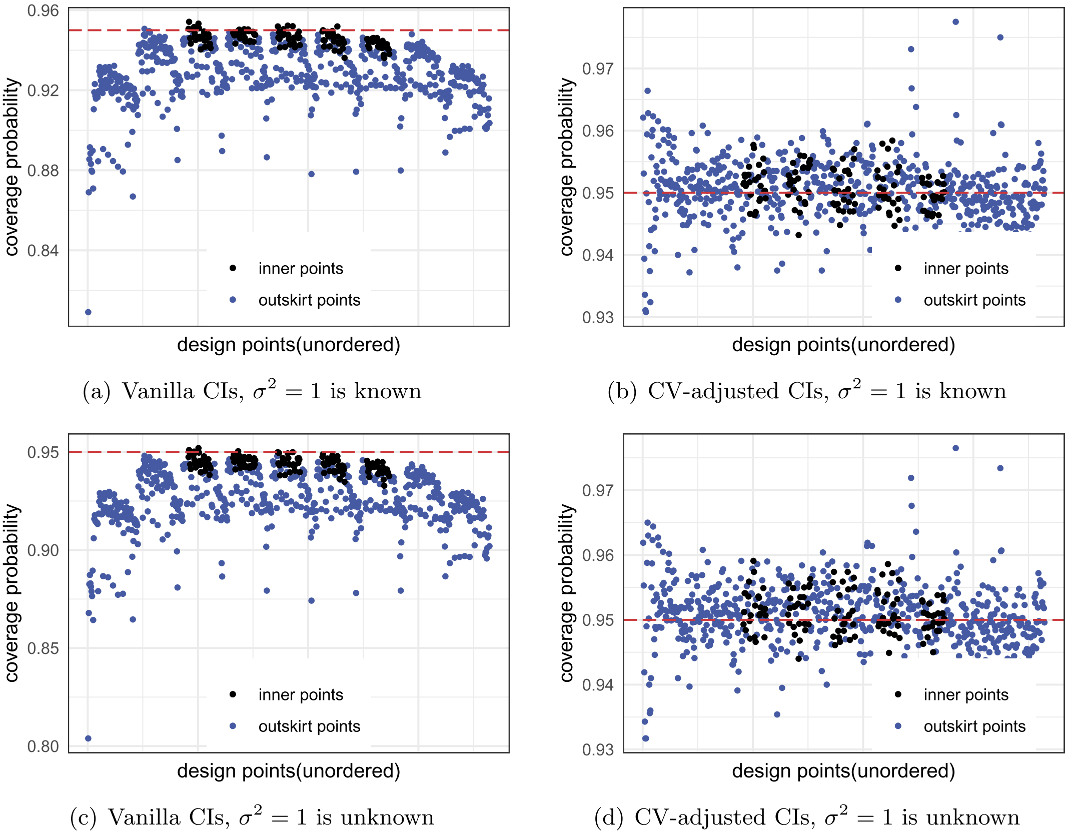


FIG. 4. *Boxplots for the coverage probabilities of the* 95% *CIs at all design points in* $d = 2$ *under known and unknown variance* $\sigma^2 = 1$, *where* $f_0(x) = e^{x_1+x_2}$ *on a* 25×25 *lattice.*



(a) Vanilla CIs, $\sigma^2 = 1$ is known

(b) CV-adjusted CIs, $\sigma^2 = 1$ is known

(c) Vanilla CIs, $\sigma^2 = 1$ is unknown

(d) CV-adjusted CIs, $\sigma^2 = 1$ is unknown

FIG. 5. *Scatter plots for the coverage probabilities of the* 95% *CIs in* $d = 3$, *where* $f_0(x) = e^{2(x_1+x_2+x_3)/3}$ *on a* $9 \times 9 \times 9$ *lattice.*

of the oracle CIs in (4.2) are given in red dashed lines. It clearly shows that the proposed CI indeed shrinks at the oracle length.

4.3. *The merit of using the block average estimator.* In (1.7) and (1.8), we propose to use the block average estimator (1.5) to carry out statistical inference about $f_0(x_0)$. It is of natural interest to ask if using either the block max–min or min–max estimator alone in the proposed procedure is adequate as this would almost reduce the computational cost by half. In this subsection, we will show empirically the benefits of using (1.5). Specifically, while of the block average estimator improves upon the block max–min estimators only slightly

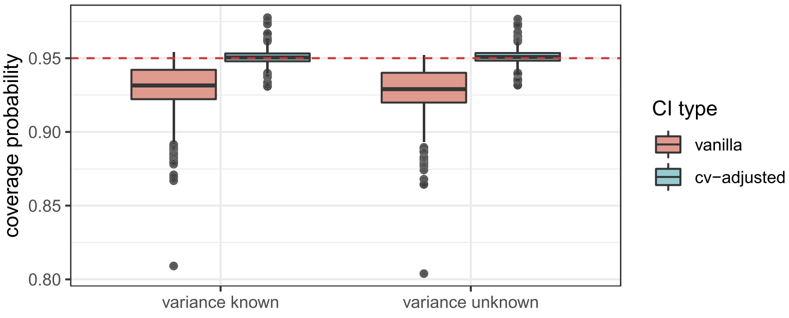


FIG. 6. *Boxplots for the coverage probabilities of the* 95% *CIs at all design points in* $d = 3$ *under known and unknown variance* $\sigma^2 = 1$, *where* $f_0(x) = e^{2(x_1+x_2+x_3)/3}$ *on a* $9 \times 9 \times 9$ *lattice.*

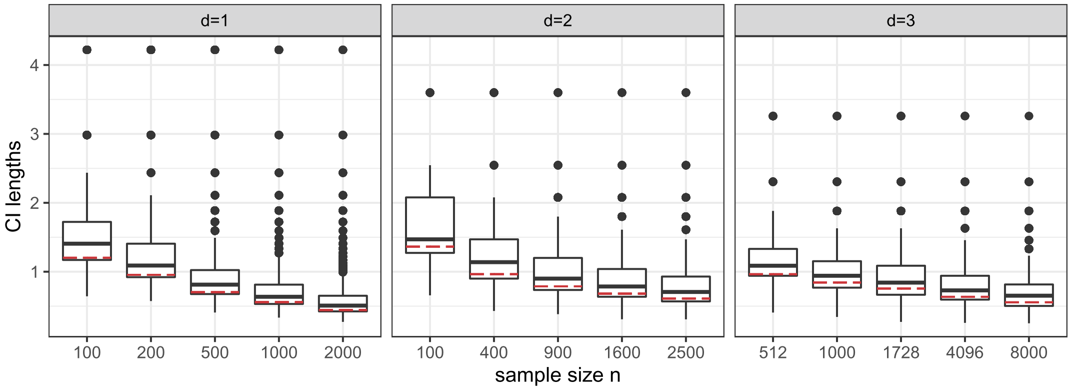


FIG. 7. *Boxplots for the lengths of the* 95% *CIs of*: (*i*) $f_0(x) = e^x$ *at* $x_0 = 0.5$ *in* $d = 1$, (*ii*) $f_0(x) = e^{x_1+x_2}$ *at* $x_0 = (0.5, 0.5)$ *in* $d = 2$ *and* (*iii*) $f_0(x) = e^{2(x_1+x_2+x_3)/3}$ *at* $x_0 = (0.5, 0.5, 0.5)$ *in* $d = 3$ *under different sample sizes* n. *Red dashed lines represent the lengths of the oracle CIs. Here* $\sigma = 1$ *is known.*

in the mean squared error (numerical results omitted), the proposed CI (1.7) which uses the block average estimator outperforms the CIs that uses the block max–min estimator alone by a fairly considerable amount in terms of accuracy for the coverage of the CIs (the situation for block min–max estimator is analogous).

More precisely, for inference based on the block max–min estimator alone, we consider the following form of the CIs:

$$\mathcal{I}_n^-(c_\delta^-) \equiv \Big[\widehat{f}_n^-(x_0) \pm c_\delta^- \cdot \widehat{\sigma}/\sqrt{n_{\widehat{u},\widehat{v}}^-(x_0)}\Big], \tag{4.3}$$

where $n_{\widehat{u},\widehat{v}}^-(x_0) \equiv \sum_i \mathbf{1}_{X_i \in [\widehat{u}^-(x_0), \widehat{v}^-(x_0)]}$ with $\widehat{u}^-(x_0), \widehat{v}^-(x_0)$ defined as any pair such that

$$\widehat{f}_n^-(x_0) \equiv \max_{u \le x_0} \min_{\substack{v \ge x_0 \\ [u,v]\cap\{X_i\}\neq\varnothing}} \bar{Y}|_{[u,v]} = \bar{Y}|_{[\widehat{u}^-(x_0), \widehat{v}^-(x_0)]}.$$

We conjecture that

CONJECTURE 1. *Under the same settings as in Theorem* 1, *for some finite random variable* $\mathbb{L}_{\boldsymbol{\alpha}}^-$ (*that does not depend on* $K(f_0, x_0)$),

$$\sqrt{n_{\widehat{u},\widehat{v}}^-(x_0)}\big(\widehat{f}_n^-(x_0) - f_0(x_0)\big) \rightsquigarrow \sigma \cdot \mathbb{L}_{\boldsymbol{\alpha}}^-.$$

Note that in $d = 1$ the block max–min and min–max estimators are equivalent, so the above conjecture reduces to Theorem 1. Below we consider $d = 2, 3$, and provide some numerical evidence that the CIs using the block average estimator could provide better probability coverage than using only the block max–min estimator based on Conjecture 1.

The summary of statistics for the coverage probabilities of the 95% CIs in $d = 2$ is listed in Table 5. The mean squared errors of the coverage probabilities of the 95% CIs by the block average estimator are 1.67×10^{-4} under known σ^2, and 1.84×10^{-4} under unknown σ^2 for vanilla CIs, reducing about 18% of those by the block max–min estimator which are 2.05×10^{-4} and 2.23×10^{-4} respectively. Similar conclusion can be made for CV-adjusted CIs.

The summary of statistics for the coverage probabilities of the 95% CIs in $d = 3$ is listed in Table 6. As the simulated critical values used in this simulation are less accurate due to the relatively small sample size in $d = 3$, the vanilla CIs for both the block average estimator and the block max–min estimator suffer from slight under-coverage. However, when using CV-adjusted CIs with improved accuracy in the mean of the coverage probabilities, similar

TABLE 5
Summary of statistics for the coverage probabilities of the 95% *CIs in* $d = 2$ *by the block average and the block max–min estimators*

	vanilla CIs				CV-adjusted CIs			
	σ known		σ u nknown		σ known		σ unknown	
	average	max–min	average	max–min	average	max–min	average	max–min
mean	0.9414	0.9405	0.9405	0.9397	0.9503	0.9503	0.9503	0.9503
median	0.9444	0.9440	0.9438	0.9432	0.9503	0.9503	0.9503	0.9503
s.d.	0.00961	0.01072	0.00973	0.01080	0.00146	0.00156	0.00140	0.00155

reduction in the mean squared errors of the coverage probabilities can be observed for the block average estimators.

4.4. *Numerical comparison with Banerjee–Wellner* (*BW*) *CIs in* $d = 1$. In this subsection, we compare numerically the coverage probabilities and the lengths of the CIs in [6] (cf. Section 2.4), hereafter referred as *BW CIs*, with the CIs proposed in this paper, labelled as *DHZ* in the plots.

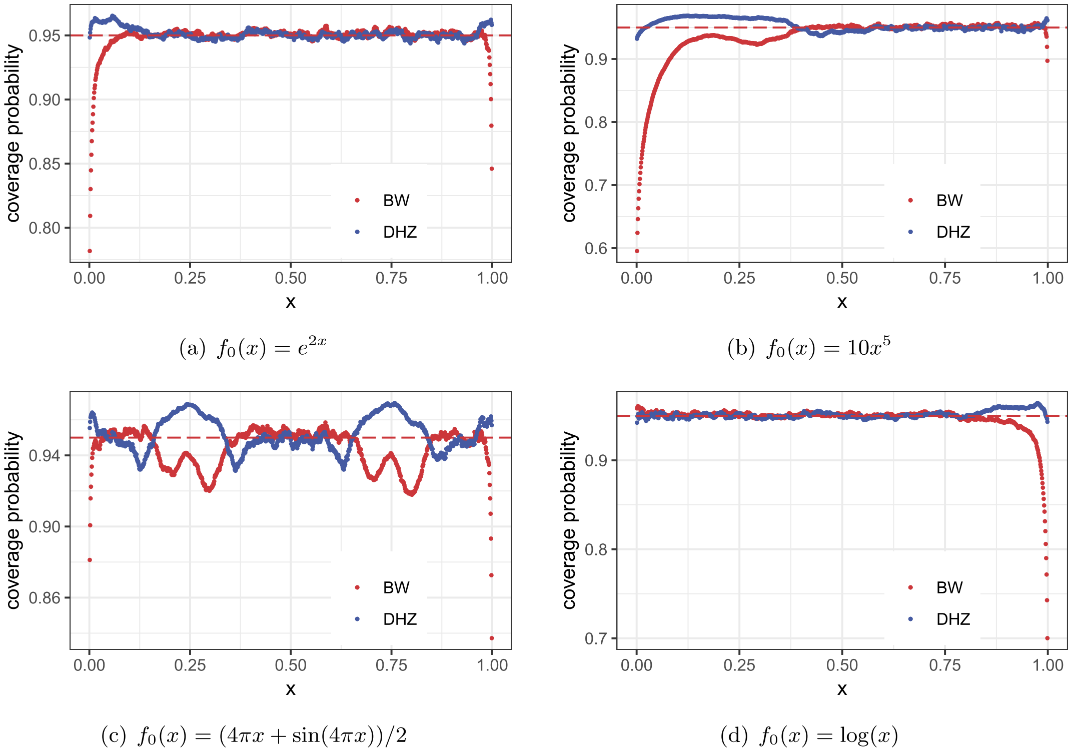
4.4.1. *Coverage probability.* For the performance on coverage, we consider four different mean functions $f_0(x)$'s: e^{2x}, $10x^5$, $(4\pi x + \sin(4\pi x))/2$ and $\log(x + 0.001)$ on $[0, 1]$. We let $n = 10^3$ and $x \in \{1/n, \dots, (n-1)/n, n/n\}$, and assume $\sigma = 1$ is known. For fair comparison, we use the vanilla CIs with the universal critical value 2.11 for the proposed CIs, and the recommended critical value 2.26916 in [6], Method 2, Table 2, for the BW CIs, both for 95% coverage. The scatter plots and boxplots for the coverage probabilities at design points $\{1/n, \dots, (n-1)/n\}$ are given in Figure 8 and Figure 9. The coverage probabilities are approximated by the relative frequencies of the successful coverage of the corresponding CIs out of $B = 10^4$ simulations.

Both types of CIs have rather accurate coverage probabilities at design points that are far from the outskirts for functions with nonextreme derivatives; see Figure 8(a) and Figure 8(d). Nevertheless, the proposed CIs appear to have two advantages. First, the edge effect in the BW CIs is much more severe than in the proposed CIs; many more outskirt points suffer under-coverage for the BW CIs. A similar phenomenon is observed in the related current status model in [30], Figure 9.16, page 270. Note that as the LSE is probably inconsistent near the edge, we do not expect the BW or the proposed CIs to provide accurate coverage in theory. However, it turns out the proposed CIs are able to give some reasonably good coverage for outskirt points numerically.

Second, the coverage probabilities of the proposed CIs over flat regions (where derivatives of the mean functions are small) are more biased towards over-coverage, while the BW CIs

TABLE 6
Summary of statistics for the coverage probabilities of the 95% *CIs in* $d = 3$ *by the block average and the block max–min estimators*

	vanilla CIs				CV-adjusted CIs			
	σ known		σ u nknown		σ known		σ unknown	
	average	max–min	average	max–min	average	max–min	average	max–min
mean	0.9273	0.9287	0.9232	0.9246	0.9513	0.9510	0.9512	0.9511
median	0.9277	0.9293	0.9236	0.9247	0.9516	0.9513	0.9515	0.9515
s.d.	0.01521	0.01450	0.01563	0.01497	0.00342	0.00428	0.00334	0.00402



(a) $f_0(x) = e^{2x}$

(b) $f_0(x) = 10x^5$

(c) $f_0(x) = (4\pi x + \sin(4\pi x))/2$

(d) $f_0(x) = \log(x)$

FIG. 8. *Scatter plots for the coverage probabilities of the* 95% *BW CIs and proposed CIs.*

are likely to suffer again under-coverage; see Figure 8(b) when $x \in [0.1, 0.5]$, and Figure 8(c) when x is around 0.25 and 0.75.

4.4.2. *CI lengths.* We compare the lengths of the BW CIs and the proposed CIs.

We may first have a glance at the BW and the proposed CIs. Let $f_0(x) = e^{2x}$ and continue the above setting in Section 4.4.1 but with $n = 10^2$. For one observation $\{(X_i, Y_i), X_i = i/n, 1 \leq i \leq n\}$, we compute both CIs for the function values at design points $\{1/n, \dots, (n-1)/n\}$ and plot them in Figure 10.

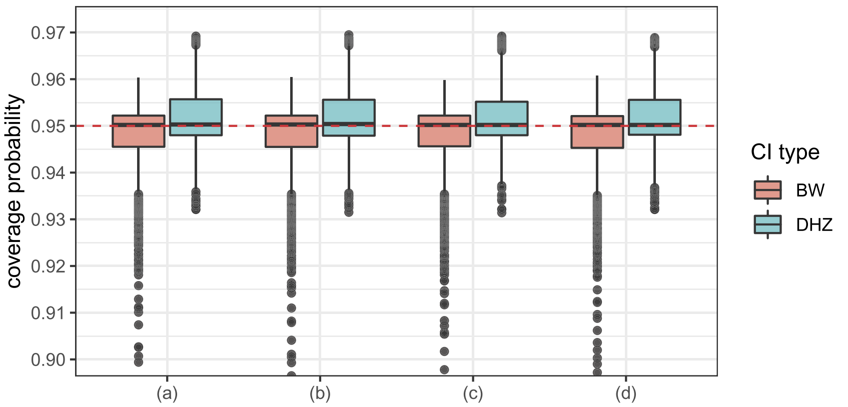
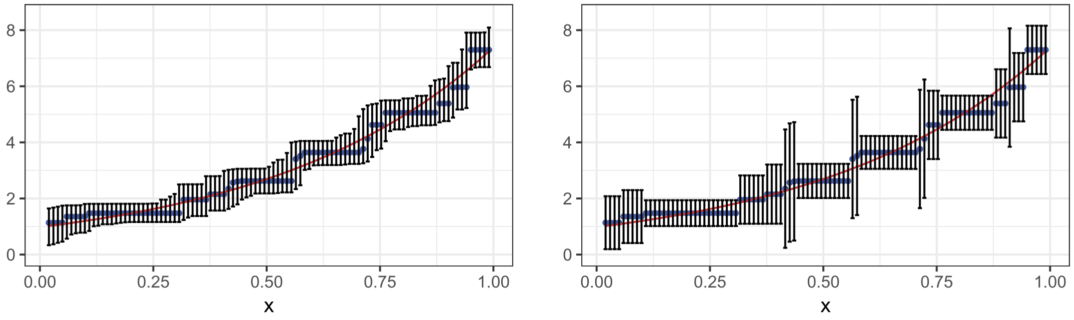


FIG. 9. *Boxplots for the coverage probabilities of the* 95% *BW CIs and proposed CIs.* (*a*), (*b*), (*c*) *and* (*d*) *correspond to the functions in Figure* 8. *Here some outliers of the boxplots for BW CI are removed as they can be as small as* 0.6.



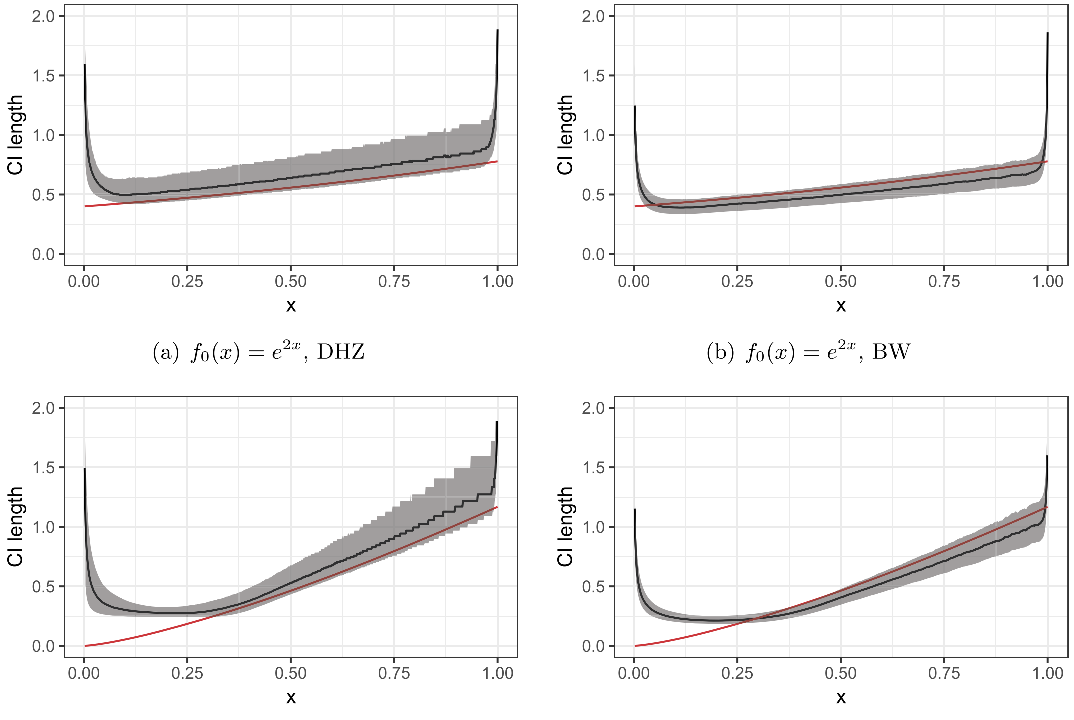
(a) BW CI

(b) proposed CI (DHZ)

FIG. 10. *95% BW CIs and proposed CIs for $f_0(x) = e^{2x}$ at $\{0.01, \dots, 0.99\}$. Here $n = 10^2$ and $\sigma = 1$ is known. Red line represents $f_0(x) = e^{2x}$ and blue dots are $\widehat{f}_n(X_i)$.*

From Figure 10, we notice an interesting difference between the BW and the proposed CIs: The lower and upper boundaries of the BW CIs seem to be monotone, but, because it is possible to have small $n_{\widehat{u},\widehat{v}}(x_0)$ for any x_0, the proposed CIs at certain locations could be quite large. We may employ some practical remedies for the proposed CI, for example, adding an extra size constraint $n_{\widehat{u},\widehat{v}}(x_0) \geq 5$ in the maximization and minimization of (1.4), but the improvement may not be as substantial as in Figure 10(a).

We also observe in this simulation that the BW CIs are narrower in Figure 10. To investigate this phenomenon more carefully, we continue the setting in Section 4.4.1 with $n = 10^3$ and compute the lengths of both CIs for design points $\{1/n, \dots, (n-1)/n\}$. We run 10^4 simulations, so that, at each design point, we obtain 10^4 BW CIs and the proposed CIs. The lengths of the CIs for $f_0(x) = e^{2x}$ and $f_0(x) = 10x^5$ are given in Figure 11. In each subfigure



(a) $f_0(x) = e^{2x}$, DHZ

(b) $f_0(x) = e^{2x}$, BW

(c) $f_0(x) = 10x^5$, DHZ

(d) $f_0(x) = 10x^5$, BW

FIG. 11. *Lengths of the 95% BW CIs and proposed CIs.*

of Figure 11, the lower (resp. upper) boundary of the shaded area represents the 1st (resp. 3rd) quantile of the lengths of 10^4 CIs, the black line is the median of the lengths, and the red line represents the length of the oracle CI defined in (4.2). Although the lengths of the proposed CI shrink at the oracle rate, which supports Theorem 2 (2.5) in $d = 1$, it seems that the BW CIs based on LRT are usually narrower than the proposed CIs.

Acknowledgments. The authors would like to thank three referees for their helpful comments and suggestions that improved the quality of the paper.

Funding. The research of H. Deng was supported in part by NSF DMS-1454817. The research of Q. Han was supported in part by NSF DMS-1916221. The research of C.-H. Zhang was supported in part by NSF DMS-1513378, DMS-1721495, IIS-1741390 and CCF-1934924.

SUPPLEMENTARY MATERIAL

Supplement: Proofs (DOI: 10.1214/20-AOS2025SUPP; .pdf). In the supplement, we provide proofs for the results in this paper.

REFERENCES

[1] BALABDAOUI, F., RUFIBACH, K. and WELLNER, J. A. (2009). Limit distribution theory for maximum likelihood estimation of a log-concave density. *Ann. Statist.* **37** 1299–1331. MR2509075 https://doi.org/10.1214/08-AOS609

[2] BALABDAOUI, F. and WELLNER, J. A. (2007). Estimation of a k-monotone density: Limit distribution theory and the spline connection. *Ann. Statist.* **35** 2536–2564. MR2382657 https://doi.org/10.1214/009053607000000262

[3] BANERJEE, M. (2007). Likelihood based inference for monotone response models. *Ann. Statist.* **35** 931–956. MR2341693 https://doi.org/10.1214/009053606000001578

[4] BANERJEE, M. (2008). Estimating monotone, unimodal and U-shaped failure rates using asymptotic pivots. *Statist. Sinica* **18** 467–492. MR2411614

[5] BANERJEE, M. and MCKEAGUE, I. W. (2007). Confidence sets for split points in decision trees. *Ann. Statist.* **35** 543–574. MR2336859 https://doi.org/10.1214/009053606000001415

[6] BANERJEE, M. and WELLNER, J. A. (2001). Likelihood ratio tests for monotone functions. *Ann. Statist.* **29** 1699–1731. MR1891743 https://doi.org/10.1214/aos/1015345959

[7] BARLOW, R. E., BARTHOLOMEW, D. J., BREMNER, J. M. and BRUNK, H. D. (1972). *Statistical Inference Under Order Restrictions. The Theory and Application of Isotonic Regression.* Wiley, London-Sydney. Wiley Series in Probability and Mathematical Statistics. MR0326887

[8] BELLEC, P. C. (2018). Sharp oracle inequalities for least squares estimators in shape restricted regression. *Ann. Statist.* **46** 745–780. MR3782383 https://doi.org/10.1214/17-AOS1566

[9] BRUNK, H. D. (1970). Estimation of isotonic regression. In *Nonparametric Techniques in Statistical Inference* (*Proc. Sympos., Indiana Univ., Bloomington, Ind.*, 1969) 177–197. Cambridge Univ. Press, London. MR0277070

[10] CHATTERJEE, S., GUNTUBOYINA, A. and SEN, B. (2015). On risk bounds in isotonic and other shape restricted regression problems. *Ann. Statist.* **43** 1774–1800. MR3357878 https://doi.org/10.1214/15-AOS1324

[11] CHATTERJEE, S., GUNTUBOYINA, A. and SEN, B. (2018). On matrix estimation under monotonicity constraints. *Bernoulli* **24** 1072–1100. MR3706788 https://doi.org/10.3150/16-BEJ865

[12] CHATTERJEE, S. and LAFFERTY, J. (2019). Adaptive risk bounds in unimodal regression. *Bernoulli* **25** 1–25. MR3892309 https://doi.org/10.3150/16-bej922

[13] CHEN, Y. and SAMWORTH, R. J. (2016). Generalized additive and index models with shape constraints. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 729–754. MR3534348 https://doi.org/10.1111/rssb.12137

[14] DENG, H., HAN, Q. and ZHANG, C.-H (2021). Supplement to "Confidence intervals for multiple isotonic regression and other monotone models." https://doi.org/10.1214/20-AOS2025SUPP

[15] DENG, H. and ZHANG, C.-H. (2020). Isotonic regression in multi-dimensional spaces and graphs. *Ann. Statist.* **48** 3672–3698. MR4185824 https://doi.org/10.1214/20-AOS1947

[16] DOSS, C. R. (2019). Concave regression: Value-constrained estimation and likelihood ratio-based inference. *Math. Program.* **174** 5–39. MR3935071 https://doi.org/10.1007/s10107-018-1338-5

[17] DOSS, C. R. and WELLNER, J. A. (2016). Global rates of convergence of the MLEs of log-concave and s-concave densities. *Ann. Statist.* **44** 954–981. MR3485950 https://doi.org/10.1214/15-AOS1394

[18] DOSS, C. R. and WELLNER, J. A. (2019). Inference for the mode of a log-concave density. *Ann. Statist.* **47** 2950–2976. MR3988778 https://doi.org/10.1214/18-AOS1770

[19] DÜMBGEN, L. and RUFIBACH, K. (2009). Maximum likelihood estimation of a log-concave density and its distribution function: Basic properties and uniform consistency. *Bernoulli* **15** 40–68. MR2546798 https://doi.org/10.3150/08-BEJ141

[20] DÜMBGEN, L., SAMWORTH, R. and SCHUHMACHER, D. (2011). Approximation by log-concave distributions, with applications to regression. *Ann. Statist.* **39** 702–730. MR2816336 https://doi.org/10.1214/10-AOS853

[21] FANG, B., GUNTUBOYINA, A. and SEN, B. (2019). Multivariate extensions of isotonic regression and total variation denoising via entire monotonicity and Hardy–Krause variation. Preprint. Available at arXiv:1903.01395.

[22] FENG, O. Y., GUNTUBOYINA, A., KIM, A. K. H. and SAMWORTH, R. J. (2018). Adaptation in multivariate log-concave density estimation. *Ann. Statist.* To appear. Available at arXiv:1812.11634.

[23] FOKIANOS, K., LEUCHT, A. and NEUMANN, M. H. (2020). On integrated L^1 convergence rate of an isotonic regression estimator for multivariate observations. *IEEE Trans. Inf. Theory* **66** 6389–6402. MR4173546

[24] GHOSAL, P. and SEN, B. (2017). On univariate convex regression. *Sankhya A* **79** 215–253. MR3707421 https://doi.org/10.1007/s13171-017-0104-8

[25] GRENANDER, U. (1956). On the theory of mortality measurement. II. *Skand. Aktuarietidskr.* **39** 125–153. MR0093415 https://doi.org/10.1080/03461238.1956.10414944

[26] GROENEBOOM, P. (1985). Estimating a monotone density. In *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer, Vol. II* (*Berkeley, Calif.*, 1983). *Wadsworth Statist./Probab. Ser.* 539–555. Wadsworth, Belmont, CA. MR0822052

[27] GROENEBOOM, P. (1989). Brownian motion with a parabolic drift and Airy functions. *Probab. Theory Related Fields* **81** 79–109. MR0981568 https://doi.org/10.1007/BF00343738

[28] GROENEBOOM, P. (2015). Rcpp scripts. Available at https://github.com/pietg/book/tree/master/Rcpp_scripts/.

[29] GROENEBOOM, P. and JONGBLOED, G. (1995). Isotonic estimation and rates of convergence in Wicksell's problem. *Ann. Statist.* **23** 1518–1542. MR1370294 https://doi.org/10.1214/aos/1176324310

[30] GROENEBOOM, P. and JONGBLOED, G. (2014). *Nonparametric Estimation Under Shape Constraints*: *Estimators, Algorithms and Asymptotics*. *Cambridge Series in Statistical and Probabilistic Mathematics* **38**. Cambridge Univ. Press, New York. MR3445293 https://doi.org/10.1017/CBO9781139020893

[31] GROENEBOOM, P. and JONGBLOED, G. (2015). Nonparametric confidence intervals for monotone functions. *Ann. Statist.* **43** 2019–2054. MR3375875 https://doi.org/10.1214/15-AOS1335

[32] GROENEBOOM, P., JONGBLOED, G. and WELLNER, J. A. (2001). A canonical process for estimation of convex functions: The "invelope" of integrated Brownian motion $+t^4$. *Ann. Statist.* **29** 1620–1652. MR1891741 https://doi.org/10.1214/aos/1015345957

[33] GROENEBOOM, P., JONGBLOED, G. and WELLNER, J. A. (2001). Estimation of a convex function: Characterizations and asymptotic theory. *Ann. Statist.* **29** 1653–1698. MR1891742 https://doi.org/10.1214/aos/1015345958

[34] GROENEBOOM, P. and WELLNER, J. A. (1992). *Information Bounds and Nonparametric Maximum Likelihood Estimation*. *DMV Seminar* **19**. Birkhäuser, Basel. MR1180321 https://doi.org/10.1007/978-3-0348-8621-5

[35] GUNTUBOYINA, A. and SEN, B. (2015). Global risk bounds and adaptation in univariate convex regression. *Probab. Theory Related Fields* **163** 379–411. MR3405621 https://doi.org/10.1007/s00440-014-0595-3

[36] HALL, P., KAY, J. W. and TITTERINGTON, D. M. (1991). On estimation of noise variance in two-dimensional signal processing. *Adv. in Appl. Probab.* **23** 476–495.

[37] HAN, Q. (2019). Global empirical risk minimizers with "shape constraints" are rate optimal in general dimensions. Preprint. Available at arXiv:1905.12823.

[38] HAN, Q. and KATO, K. (2019). Berry–Esseen bounds for Chernoff-type non-standard asymptotics in isotonic regression. Preprint. Available at arXiv:1910.09662.

[39] HAN, Q., WANG, T., CHATTERJEE, S. and SAMWORTH, R. J. (2019). Isotonic regression in general dimensions. *Ann. Statist.* **47** 2440–2471. MR3988762 https://doi.org/10.1214/18-AOS1753

[40] HAN, Q. and WELLNER, J. A. (2016). Approximation and estimation of s-concave densities via Rényi divergences. *Ann. Statist.* **44** 1332–1359. MR3485962 https://doi.org/10.1214/15-AOS1408

[41] HAN, Q. and WELLNER, J. A. (2016). Multivariate convex regression: Global risk bounds and adaptation. Preprint. Available at arXiv:1601.06844.

[42] HAN, Q. and ZHANG, C.-H. (2020). Limit distribution theory for block estimators in multiple isotonic regression. *Ann. Statist.* **48** 3251–3282. MR4185808 https://doi.org/10.1214/19-AOS1928

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

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

[45] JANKOWSKI, H. K. and WELLNER, J. A. (2009). Nonparametric estimation of a convex bathtub-shaped hazard function. *Bernoulli* **15** 1010–1035. MR2597581 https://doi.org/10.3150/09-BEJ202

[46] KIM, A. K. H., GUNTUBOYINA, A. and SAMWORTH, R. J. (2018). Adaptation in log-concave density estimation. *Ann. Statist.* **46** 2279–2306. MR3845018 https://doi.org/10.1214/17-AOS1619

[47] KIM, A. K. H. and SAMWORTH, R. J. (2016). Global rates of convergence in log-concave density estimation. *Ann. Statist.* **44** 2756–2779. MR3576560 https://doi.org/10.1214/16-AOS1480

[48] KOENKER, R. and MIZERA, I. (2010). Quasi-concave density estimation. *Ann. Statist.* **38** 2998–3027. MR2722462 https://doi.org/10.1214/10-AOS814

[49] KOSOROK, M. R. (2008). Bootstrapping in Grenander estimator. In *Beyond Parametrics in Interdisciplinary Research: Festschrift in Honor of Professor Pranab K. Sen. Inst. Math. Stat.* (*IMS*) *Collect.* **1** 282–292. IMS, Beachwood, OH. MR2462212 https://doi.org/10.1214/193940307000000202

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

[51] LIM, E. and GLYNN, P. W. (2012). Consistency of multidimensional convex regression. *Oper. Res.* **60** 196–208. MR2911667 https://doi.org/10.1287/opre.1110.1007

[52] MAMMEN, E. (1991). Nonparametric regression under qualitative smoothness assumptions. *Ann. Statist.* **19** 741–759. MR1105842 https://doi.org/10.1214/aos/1176348118

[53] MAMMEN, E., MARRON, J. S., TURLACH, B. A. and WAND, M. P. (2001). A general projection framework for constrained smoothing. *Statist. Sci.* **16** 232–248. MR1874153 https://doi.org/10.1214/ss/1009213727

[54] MUNK, A., BISSANTZ, N., WAGNER, T. and FREITAG, G. (2005). On difference-based variance estimation in nonparametric regression when the covariate is high dimensional. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 19–41. MR2136637 https://doi.org/10.1111/j.1467-9868.2005.00486.x

[55] PRAKASA RAO, B. L. S. (1969). Estimation of a unimodal density. *Sankhyā Ser. A* **31** 23–36. MR0267677

[56] PRAKASA RAO, B. L. S. (1970). Estimation for distributions with monotone failure rate. *Ann. Math. Stat.* **41** 507–519. MR0260133 https://doi.org/10.1214/aoms/1177697091

[57] RICE, J. (1984). Bandwidth choice for nonparametric regression. *Ann. Statist.* **12** 1215–1230. MR0760684 https://doi.org/10.1214/aos/1176346788

[58] ROBERTSON, T., WRIGHT, F. T. and DYKSTRA, R. L. (1988). *Order Restricted Statistical Inference. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics*. Wiley, Chichester. MR0961262

[59] SEIJO, E. and SEN, B. (2011). Change-point in stochastic design regression and the bootstrap. *Ann. Statist.* **39** 1580–1607. MR2850213 https://doi.org/10.1214/11-AOS874

[60] SEIJO, E. and SEN, B. (2011). Nonparametric least squares estimation of a multivariate convex regression function. *Ann. Statist.* **39** 1633–1657. MR2850215 https://doi.org/10.1214/10-AOS852

[61] SEN, B. and BANERJEE, M. (2007). A pseudolikelihood method for analyzing interval censored data. *Biometrika* **94** 71–86. MR2307901 https://doi.org/10.1093/biomet/asm011

[62] SEN, B., BANERJEE, M. and WOODROOFE, M. (2010). Inconsistency of bootstrap: The Grenander estimator. *Ann. Statist.* **38** 1953–1977. MR2676880 https://doi.org/10.1214/09-AOS777

[63] SEREGIN, A. and WELLNER, J. A. (2010). Nonparametric estimation of multivariate convex-transformed densities. *Ann. Statist.* **38** 3751–3781. MR2766867 https://doi.org/10.1214/10-AOS840

[64] SUN, J. and KALBFLEISCH, J. D. (1995). Estimation of the mean function of point processes based on panel count data. *Statist. Sinica* **5** 279–289. MR1329298

[65] VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes. Springer Series in Statistics*. Springer, New York. MR1385671 https://doi.org/10.1007/978-1-4757-2545-2

[66] WELLNER, J. A. and ZHANG, Y. (2000). Two estimators of the mean of a counting process with panel count data. *Ann. Statist.* **28** 779–814. MR1792787 https://doi.org/10.1214/aos/1015951998

[67] WRIGHT, F. T. (1981). The asymptotic behavior of monotone regression estimates. *Ann. Statist.* **9** 443–448. MR0606630

[68] XU, M. and SAMWORTH, R. J. (2021). High-dimensional nonparametric density estimation via symmetry and shape constraints. *Ann. Statist.* To appear. Available at arXiv:1903.06092.

[69] ZHANG, C.-H. (2002). Risk bounds in isotonic regression. *Ann. Statist.* **30** 528–555. MR1902898 https://doi.org/10.1214/aos/1021379864