Submitted to the Annals of Statistics
arXiv: arXiv:0000.0000

DIVIDE AND CONQUER IN NON-STANDARD PROBLEMS AND THE SUPER-EFFICIENCY PHENOMENON

By Moulinath Banerjee[*], Cécile Durot[†] and Bodhisattva Sen[‡]

University of Michigan, Université Paris Nanterre and Columbia University

We study how the divide and conquer principle works in non-standard problems where rates of convergence are typically slower than $\sqrt{n}$ and limit distributions are non-Gaussian, and provide a detailed treatment for a variety of important and well-studied problems involving nonparametric estimation of a monotone function. We find that the pooled estimator, obtained by averaging non-standard estimates across mutually exclusive subsamples, outperforms the non-standard monotonicity-constrained (global) estimator based on the entire sample in the sense of *pointwise inference*. We also show that, under appropriate conditions, if the number of subsamples is allowed to increase at appropriate rates, the pooled estimator is asymptotically normally distributed with a variance that is empirically estimable from the subsample-level estimates. Further, in the context of monotone regression, we show that this gain in pointwise efficiency comes at a price — the pooled estimator's performance, in a *uniform sense* (maximal risk) over a class of models worsens as the number of subsamples increases, leading to a version of the super-efficiency phenomenon. In the process, we develop analytical results for the order of the bias in isotonic regression, which are of independent interest.

1. Introduction.

Suppose that $W_1, \ldots, W_N$ are i.i.d. random elements having a common distribution P. We assume that P is unknown and $\theta_0 \equiv \theta_0(P)$ is a finite dimensional parameter of interest. In this paper we focus on non-standard statistical problems where a natural estimator $\hat{\theta}$ (of θ_0) converges in distribution to a non-normal limit at a rate slower than $N^{1/2}$, i.e.,

$$r_N(\hat{\theta} - \theta_0) \overset{d}{\to} G, \tag{1.1}$$

where $r_N = o(\sqrt{N})$ and G is non-normal, has mean zero and finite variance σ^2. However, σ^2 can depend on P in a complicated fashion which often makes it difficult to use (1.1) to construct confidence intervals (CIs) and hypothesis tests for θ_0. Such non-standard limits primarily arise due to the inherent lack of smoothness in the underlying estimation procedure. Also, in many such scenarios the computation of $\hat{\theta}$ is complicated, requiring computationally intensive algorithms. Thus, in the face of a humongous sample size N — quite common with present-day data — these problems present a significant challenge both in computation and inference.

[*] Supported by NSF Grant DMS-1308890

[†] Supported by Labex MME-DII (ANR11-LBX-0023-01)

[‡] Supported by NSF CAREER Grant DMS-1150435

AMS 2000 subject classifications: Primary 62G20, 62G08; secondary 62F30

Keywords and phrases: cube-root asymptotics, isotonic regression, local minimax risk, non-Gaussian limit, sample-splitting

In this paper, we investigate how such non-standard estimates behave under the "divide-and-conquer" strategy – a method that has been much used in the analysis of massive data sets; see e.g., [20, 29, 30] – with an emphasis on function estimation under monotonicity constraints which constitutes an important genre of non-standard problems of the above type. Indeed, a rich class of non-standard problems arises in the nonparametric maximum likelihood/least-squares (NPMLE/LSE) based estimation of a monotone function, an important sub-area of the field known as shape-restricted inference which has seen much activity over the last few decades. The literature on monotone function estimation and inference is extensive: for an excellent exposition, we direct the reader to the recent text [15]. A key feature of the NPMLE/LSE of a monotone function under standard smoothness assumptions is the pointwise $n^{1/3}$ rate of convergence to the truth with a non-Gaussian mean 0 limit distribution. Such estimators have been studied in a variety of interesting statistical contexts, e.g., isotonic regression [7, 8], where a monotone regression function is estimated via least squares under that shape constraint, the current status model (and extensions thereof) [5, 16], where the distribution of a failure time is estimated under the monotonicity constraint from discrete response data, Grenander's problem of estimating a decreasing density [14, 23], nonparametric estimation of a monotone failure rate [3, 18], likelihood based inference for monotone response models [2], to name a few.

To provide a glimpse of the asymptotic features in monotone function estimation, we elaborate on the first of the aforementioned examples: the isotonic regression problem. Consider i.i.d. data $\{W_i := (X_i, Y_i) : i = 1, \ldots, N\}$ from the regression model $Y = \mu(X) + \epsilon$ where $Y \in \mathbb{R}$ is the response variable, $X \in [0,1]$ (with density f) is the covariate, μ is the unknown *nonincreasing* regression function, $\mathbb{E}(\epsilon|X) = 0$, and the conditional variance $v^2(x) := \mathbb{E}(\epsilon^2|X = x)$ is finite. The goal is to estimate $\mu : [0,1] \to \mathbb{R}$ nonparametrically, under the constraint of monotonicity. We will consider the LSE $\hat{\mu}$ defined as a minimizer of $\psi \mapsto \sum_{i=1}^n (Y_i - \psi(X_i))^2$ over the set of all nonincreasing functions $\psi : [0,1] \to \mathbb{R}$. We know that $\hat{\mu}$ is unique at the data points X_i's and is connected to the slope of the least concave majorant of the cumulative sum diagram [25, Chapter 1]. If $\mu'(t_0) \neq 0$, where t_0 is an interior point in the support of X, and v is continuous,

$$N^{1/3}(\hat{\mu}(t_0) - \mu(t_0)) \xrightarrow{d} \kappa\mathbb{Z}, \tag{1.2}$$

with $\kappa := |4v(t_0)^2\mu'(t_0)/f(t_0)|^{1/3}$ and $\mathbb{Z} := \operatorname{argmin}_{s\in\mathbb{R}}\{W(s) + s^2\}$ (where W is a standard two-sided Brownian motion starting at 0) has the so-called Chernoff's distribution; see [28, Theorem 1]. It is known that $\mathbb{Z}$ is symmetric (around 0) and has mean zero. Lastly $\sigma^2 = \mathrm{Var}(\kappa\mathbb{Z})$, the variance of the limiting distribution, is difficult to estimate as it involves the derivative of μ, the estimation of which is well-known to be a challenging problem [5].

A closely related problem is the estimation of the inverse isotonic function at a point. If a is an interior point in the range of μ and $t_0 = \mu^{-1}(a) \in (0,1)$ satisfies $\mu'(t_0) \neq 0$, then

$$N^{1/3}(\hat{\mu}^{-1}(a) - \mu^{-1}(a)) \xrightarrow{d} \tilde{\kappa}\mathbb{Z}, \tag{1.3}$$

where $\tilde{\kappa} := |4v^2(t_0)/\mu'(t_0)^2 f(t_0)|^{1/3}$; this can be derived, e.g., from the arguments in [11]. Similar results hold across a vast array of monotone function problems.

Another class of problems sharing the same convergence rate and exhibiting non-standard behavior is found in the world of "cube-root asymptotics" [19], which include, e.g., estimation of the mode [9], Manski's maximum score estimator [21], change-point estimation under smooth mis-specification [4], least absolute median of squares [26], shorth estimation [13].

Divide and Conquer/Sample splitting: In the *sample-splitting* strategy called *divide-and-conquer*, the available data is partitioned into subsamples, an estimate of θ_0 is computed from each subsample, and finally the subsample level estimates are combined appropriately to form the final estimator. Our combining/pooling strategy will be based on averaging. To be precise, assume that N is large and write $N = n \times m$, where n is still large and m relatively smaller (e.g., $n = 1000$, $m = 50$, so that $N = 50000$). We define our new "averaged" estimator as follows:

1. Divide the set of samples $W_1, \ldots, W_N$ into m disjoint subsets $S_1, \ldots, S_m$.
2. For each $j = 1, \ldots, m$, compute the estimator $\hat{\theta}_j$ based on the data points in S_j.
3. Average together these estimators to obtain the final 'pooled' estimator:

$$\bar{\theta} = \frac{1}{m} \sum_{j=1}^{m} \hat{\theta}_j. \tag{1.4}$$

Observe that if the computation of $\hat{\theta}$, the global estimator based on all N observations, is of super-linear computational complexity in the sample size, computing $\bar{\theta}$ saves resources compared to $\hat{\theta}$. Further, the computation of $\bar{\theta}$ can be readily parallelized, using m CPU's. Such averaged estimators have been considered by many authors recently to estimate nonparametric functions, but typically under smoothness constraints; see e.g., [29, 30], and also [20] for a discussion with a broader scope. The above papers illustrate that the approach significantly reduces the required amount of primary memory and computation time in a variety of cases, yet statistical optimality — in the sense that the resulting estimator is as efficient as the global estimate — is retained.

We next lay down the contributions of our paper to the divide and conquer literature.

1. In Sections 2 and 3, we present general results on the asymptotic distribution of the averaged (pooled) estimator $\bar{\theta}$, both when m is fixed and when allowed to increase as N increases, in which case a normal distribution arises in the limit (Theorem 3.1). Furthermore, in the latter case, allowable choices of m, which affect the rate of convergence of $\bar{\theta}$, crucially depend on the order of the bias of $\hat{\theta}_j$. Pooling provides us with a novel way to construct a CI for θ_0 whose length is shorter than that of using $\hat{\theta}$ owing to the faster convergence rate involved. The calibration of the new CI involves normal quantiles, instead of quantiles of the non-standard limits that describe $\hat{\theta}$ asymptotically. Moreover, the variance σ^2 can be estimated empirically using the subsample-level estimates, whereas in the method involving $\hat{\theta}$, one is typically forced to impute values of several nuisance parameters that arise in the expression for σ^2 using estimates that can be quite unreliable.

2. The possible gain by sample-splitting is driven by the bias of the non-standard estimator, that one needs to quantify. In Section 4, we provide results on the bias of monotonicity constrained estimators as well as their inverses in a variety of important nonparametric problems : isotonic regression, current status (case 1 interval) censoring, Grenander's decreasing density problem, and the problem of estimating a monotone failure rate. The bias in these problems is *hard* to compute because the usual Taylor expansion arguments that work in smooth function estimation problems cannot be employed. For the first time, we provide a non-trivial bound on the order of the bias of the monotone LSE/NPMLE under mild regularity assumptions.

Furthermore, establishing the asymptotic normality of the pooled estimator in monotone function problems requires showing uniform integrability of certain powers of the normalized LSE/NPMLE as well as its inverse, pointwise. We establish this property for all powers $p \geq 1$ under suitable model-specific assumptions in Section 4. As a consequence, we obtain upper bounds on the maximal risk of the monotone LSE/NPMLE and its inverse over suitable classes of functions. Although such bounds on the maximal risk are known for most nonparametric function estimators, this is the first instance of such a result in the general isotonic regression problem[1]. The results on bias and uniform integrability are then used to study the sample-splitting method in the different models considered, by verifying the conditions of Theorem 3.1.

3. In Section 5, we present a rigorous study of a super-efficiency phenomenon that comes into play when using the pooled estimator in the context of estimating the inverse of an isotonic regression function. Let $\overline{\theta}$ denote the average of the $\hat{\mu}_{n,j}^{-1}(a)$'s, where $\hat{\mu}_{n,j}$ is the isotonic LSE from the j'th subsample and let $\theta_0 := t_0 \equiv \mu^{-1}(a)$ (see (1.3)). We show that for a suitably chosen (large enough) class of models $\mathcal{M}_0$, when $m \equiv m_n \to \infty$, the maximal risk,

$$\sup_{\mu \in \mathcal{M}_0} \mathbb{E}_\mu \left[N^{2/3} (\overline{\theta} - \theta_0)^2 \right]$$

diverges to infinity as $N \to \infty$ whereas the corresponding maximal risk of the global estimator $\hat{\theta} \equiv \hat{\mu}_N^{-1}(a)$ remains bounded. Thus, while the pooled estimator $\overline{\theta}$ can outperform the LSE under any *fixed* model, its performance over a class of models is compromised relative to the isotonic LSE. The larger the number of splits (m), the better the performance under a fixed model, but the worse the performance over the entire class. Our discoveries therefore serve as a cautionary tale that illustrates the potential pitfalls of using sample-splitting: the benefits from sample-splitting, both computational and in the sense of *pointwise inference* may come at subtle costs. The proofs of some of the main results are presented in Section 7 and the supplement provides detailed coverage of additional technical material.

We note that the class of non-standard problems is large and varied and an integrated treatment of divide and conquer *across different genres of non-standard problems* [where the core technical challenges lie in the bias calculations and uniform integrability considerations] appears infeasible, since the tools and techniques, which are non-trivial, will vary quite substantially from genre to genre. In this paper, we have developed our computations for several examples belonging to a single but important genre — namely monotone function estimation — which submit to a reasonably unified treatment, and hope that the interesting findings of this paper will spur further studies of divide and conquer in other classes of non-standard problems.

Before we move on to the rest of the paper there is one point on which some clarity needs to be provided: in subsequent sections, the total sample size N will be written as $m \times n$. Now, starting with an m, not all sample-sizes N can be represented as a product of that form. To get around this difficulty, one can work with the understanding that we reduce our sample size from N to $\tilde{N} := m \times \lfloor N/m \rfloor$ (which is then renamed N) with the last few samples being discarded. Since finitely many are discarded, the resulting pooled estimate will be as precise in an asymptotic sense as the one based on the original N, provided m is of a smaller order than N (which will always be the case in the sequel). In this paper we work with the $\tilde{N}$ interpretation.

2. Fixed m and growing n.

Consider the setup of (1.1), where θ_0 is the parameter of interest and let $\bar{\theta}$ be the pooled estimator as defined in (1.4). We start with a simple lemma that illustrates the statistical benefits of sample-splitting when n is large and m is held fixed.

[1] Similar risk-bounds are presented in the special case of current status model in Theorem 11.3 in [15]; however, their derivation uses a special feature of the isotonic estimator in that particular model which is not true in the general scenarios we consider, as discussed later in Remark 4.1.

LEMMA 2.1. *Suppose (1.1) holds with $\mathbb{E}(G) = 0$ and $\mathrm{Var}(G) = \sigma^2$. For m fixed and $N = m \times n$,*

$$\sqrt{m} r_n(\bar{\theta} - \theta_0) \xrightarrow{d} H := m^{-1/2}(G_1 + G_2 + \ldots + G_m), \qquad \text{as } n \to \infty, \tag{2.1}$$

where $G_1, G_2, \ldots, G_m$ are i.i.d. G. Note that H has mean zero and variance σ^2.

Compare the above result with the fact that if all N data points were used together to obtain $\hat{\theta}$ we would have the limiting distribution in (1.1): if $\{[r_N(\hat{\theta} - \theta_0)]^2\}_{n \geq 1}$ is uniformly integrable (which we will prove later for certain problems), we conclude that $\mathbb{E}[r_N^2(\hat{\theta} - \theta_0)^2]$ converges to $\mathrm{Var}(G)$ as $N \to \infty$, while

$$\mathbb{E}\left[\frac{m r_n^2}{r_N^2} r_N^2(\bar{\theta} - \theta_0)^2\right] \to \mathrm{Var}(G)\,, \qquad \text{as } N \to \infty, \tag{2.2}$$

since G and H have the same variance. Thus, the asymptotic relative efficiency of $\bar{\theta}$ with respect to $\hat{\theta}$ is $m r_n^2 / r_N^2$. For example, if $r_N = N^\gamma, \gamma < 1/2$, then using $\bar{\theta}$ gives us a reduction in asymptotic variance by a factor of $m^{1-2\gamma}$. Hence, for estimating θ_0, the pooled estimator $\bar{\theta}$ *outperforms* $\hat{\theta}$.

REMARK 2.1. *If $\{[r_n(\hat{\theta}_j - \theta_0)]^2\}_{n \geq 1}$ is uniformly integrable, then the variance of $r_n(\hat{\theta}_j - \theta_0)$, which is equal to $\sigma_n^2 := r_n^2 \mathrm{Var}(\hat{\theta}_j)$, converges to σ^2 as $n \to \infty$, for every $j = 1, \ldots, m$. As we have m independent replicates from the distribution of $\hat{\theta}_j$, σ^2 can be approximated by*

$$\hat{\sigma}^2 := \frac{r_n^2}{m - 1} \sum_{j=1}^{m} (\hat{\theta}_j - \bar{\theta})^2. \tag{2.3}$$

REMARK 2.2. *For moderate values of m, the m-fold convolution H in (2.1) can be approximated by an appropriate t (or normal) distribution. This yields a simple and natural way to construct an approximate $(1-\alpha)$ CI for θ_0 that completely by-passes the direct estimation of the problematic nuisance parameter σ^2:*

$$\left[\bar{\theta} - \frac{\hat{\sigma}}{r_n \sqrt{m}} t_{\alpha/2, m-1}, \bar{\theta} + \frac{\hat{\sigma}}{r_n \sqrt{m}} t_{\alpha/2, m-1}\right],$$

where $t_{\alpha, m-1}$ denotes the $(1-\alpha)$-th quantile of the t-distribution with $m-1$ degrees of freedom. Furthermore, in certain cases , it is the case that we know the distribution of the centered non-Gaussian variable $\tilde{Z} := G/\sigma$ (which has unit variance) and are able to simulate from its distribution.[2] In this case, the Student's t (or normal) approximation can be avoided. As $r_n(\hat{\theta}_j - \theta_0) \xrightarrow{d} \sigma\tilde{Z}$, for $j = 1, \ldots, m$, the asymptotic distribution of $\hat{\sigma}^{-1} r_n m^{1/2}(\bar{\theta} - \theta_0)$ coincides with the distribution of

$$\tilde{H} := \left[\sum_{i=1}^{m} (\tilde{Z}_i - \bar{Z}_m)^2/(m-1)\right]^{-1} m^{1/2} \bar{Z}_m,$$

where $\tilde{Z}_1, \ldots, \tilde{Z}_m$ are i.i.d. copies of $\tilde{Z}$, and $\bar{Z}_m$ denotes their sample mean. Hence one could replace the t-distribution with the appropriate quantiles of $\tilde{H}$, which can be computed, thanks to the fact that we are able to simulate from the distribution of $\tilde{Z}$.

[2] For example, in monotone function estimation, e.g. (1.2), $\tilde{Z}$ is the Chernoff random variable scaled by its own standard deviation.

3. Letting m grow with n: asymptotic considerations. In this section, we derive the asymptotic distribution of $\sqrt{m}r_n(\bar{\theta} - \theta_0)$ under certain conditions, as $m \to \infty$. To highlight the dependence on n, we write $m \equiv m_n$, $\hat{\theta}_j \equiv \hat{\theta}_{n,j}$ and $\bar{\theta} = \bar{\theta}_{m_n}$. Consider the triangular array $\{\xi_{n,1}, \xi_{n,2}, \ldots, \xi_{n,m_n}\}_{n \geq 1}$ where $\xi_{n,j} := r_n(\hat{\theta}_{n,j} - \theta_0)$. Let $b_n := \mathbb{E}(\xi_{n,1}) = r_n(\theta_n - \theta_0)$ where $\theta_n := \mathbb{E}(\hat{\theta}_{n,1})$ is assumed to be well-defined. The following theorem is proved in Section 7.1.

THEOREM 3.1. *Suppose (1.1) holds where $\mathbb{E}(G) = 0$ and $\mathrm{Var}(G) = \sigma^2$. Also, suppose that $b_n = O(c_n^{-1})$, where $c_n \to \infty$ as $n \to \infty$, and $\{\xi_{n,1}^2\}$ is uniformly integrable. Then, as $n \to \infty$,*

(i) for any $m_n \to \infty$ such that $m_n = o(c_n^2)$, $\sqrt{m_n}r_n(\bar{\theta}_{m_n} - \theta_0) \xrightarrow{d} N(0, \sigma^2)$;

(ii) if $m_n \sim O(c_n^2)$, and furthermore $\sqrt{m_n}\, b_n \to \tau$, then $\sqrt{m_n}r_n(\bar{\theta}_{m_n} - \theta_0) \xrightarrow{d} N(\tau, \sigma^2)$.

REMARK 3.1 (Gains from sample-splitting: "divide to conquer"). *The pooled estimator $\bar{\theta}_{m_n}$ is more effective than $\hat{\theta}_N$, when its convergence rate exceeds that of the latter, i.e.,*

$$\frac{r_N}{\sqrt{m_n}r_n} \to 0 \Leftrightarrow \frac{r_N/r_n}{m_n^{1/2}} \to 0;$$

thus, if $r_N = N^\alpha$, using $N = n \times m_n$, this requires $\alpha < 1/2$. In other words, acceleration is only possible if the initial estimator has a slower convergence rate than the parametric rate.

REMARK 3.2 (Choice of m_n). *As above, let $r_N = N^\alpha$ with $\alpha < 1/2$, and let $c_n = n^\phi$. Choosing $m_n = n^{2\phi-\delta}$, with $0 < \delta < 2\phi$, so that $m_n = o(c_n^2)$, we have $\sqrt{m_n}\, r_n = n^{\phi-\delta/2+\alpha}$. Using $m_n \times n = N$, we get $n = N^{1/(2\phi-\delta+1)}$. The convergence rate of the pooled estimator in terms of the total sample size is therefore $N^{(\phi-\delta/2+\alpha)/2(\phi-\delta/2+1/2)}$. Since $\alpha < 1/2$, this rate is strictly less than $N^{1/2}$. Next, the improvement in the convergence rate is given by*

$$\frac{\phi - \delta/2 + \alpha}{2(\phi - \delta/2 + 1/2)} - \alpha = 2\left(\frac{1}{2} - \alpha\right) \frac{\phi - \delta/2}{\phi - \delta/2 + 1/2},$$

which is monotone decreasing in δ. This means that smaller values of δ, corresponding to larger values of $m_n = N^{(2\phi-\delta)/(2\phi-\delta+1)}$ give greater improvements in the convergence rate. In the situation of conclusion (ii) of the above theorem, when $\delta = 0$ and $m_n = O(c_n^2)$, we get the maximal convergence rate: $N^{(\alpha+\phi)/2(\phi+1/2)}$. To get the best possible rate, we would like to get hold of the optimal *value of c_n, i.e., we would want $b_n = O(c_n^{-1})$ but not $o(c_n^{-1})$. The optimal c_n might, of course, be difficult to obtain in a particular application; however, sub-optimal c_n's will also improve the rate of convergence, albeit not to the best possible extent.*

From Theorem 3.1 we see that the two key challenges to establishing the asymptotic normality of the pooled estimator are: (a) establishing uniform integrability as desired above, and, (b) determining an order for the bias b_n. In the following sections we consider the example of monotone regression and address (a) and (b) for the isotonic MLE and its inverse.

4. Sample splitting in a variety of monotone function problems. In this section, we study the behavior of the pooled estimator obtained via sample-splitting through a variety of examples involving the estimation of a monotone function in different contexts: regression, current status data, density estimation, and hazard rate estimation under right censoring. These four scenarios cover the four core statistical contexts in which monotone function estimation has been studied extensively in the literature. As we will see, the results obtained in the four different scenarios (under broadly similar assumptions) show the recurrence of the same convergence rates.

4.1. *The isotonic regression problem.* Our formal treatment is developed in the framework of [11] which considers a general monotone nonincreasing regression model described below. The results, of course, extend immediately to the nondecreasing case. Having observed i.i.d. copies $\{W_i \equiv (X_i, Y_i) : i = 1, \dots, n\}$ of $(X, Y) \in [0,1] \times \mathbb{R}$, we aim at estimating the regression function μ defined by $\mu(x) = \mathbb{E}(Y|X = x)$, for $x \in [0,1]$, under the constraint that it is nonincreasing on $[0,1]$. Alternatively, we may be interested in estimating the inverse function μ^{-1}. With $\epsilon = Y - \mu(X)$ we define $v^2(x) := \mathbb{E}(\epsilon_i^2 | X_i = x)$ for all $x \in [0,1]$ and we make the following assumptions.

(R1) μ is differentiable and decreasing on $[0,1]$ with $\inf_t |\mu'(t)| > 0$ and $\sup_t |\mu'(t)| < \infty$.
(R2) X has a bounded density f which is bounded away from zero.
(R3) There exists $c_0 > 0$ such that $v^2(t) \geq c_0(t \wedge (1-t))$ for all $t \in [0,1]$.
(R4) There exists $\alpha > 0$ such that $\mathbb{E}\left(e^{\theta\epsilon}|X\right) \leq \exp(\alpha\theta^2)$ *a.e.* for all $\theta \in \mathbb{R}$.

Assumption (R3) is less restrictive than the usual assumption of a variance function v bounded away from zero and allows us to handle the current status model in Subsection 4.2. Assumption (R4) is fulfilled for instance if the conditional distribution of ϵ given X is sub-Gaussian and the variance function v^2 is bounded.

4.1.1. *The isotonic LSE of μ and the inverse estimator.* We start with an exposition of the characterization of the LSE of μ and its inverse under the monotonicity constraint. With $X_{(1)} < \cdots < X_{(n)}$ the order statistics corresponding to $X_1, \dots, X_n$, and $Y_{(i)}$ the observation corresponding to $X_{(i)}$, let Λ_n be the piecewise-linear process on $[0,1]$ such that

$$\Lambda_n\left(\frac{i}{n}\right) = \frac{1}{n}\sum_{j \leq i} Y_{(j)} \tag{4.1}$$

for all $i \in \{0, \dots, n\}$, where we set $\sum_{j \leq 0} Y_{(j)} = 0$. Let $\hat{\lambda}_n$ be the left-hand slope of the least concave majorant of Λ_n. It is well known that a monotone $\hat{\mu}_n$ is an LSE if and only if it satisfies

$$\hat{\mu}_n(X_{(i)}) = \hat{\lambda}_n(i/n) \tag{4.2}$$

for all $i = 1, \dots, n$. In the sequel, we consider the piecewise-constant left-continuous LSE $\hat{\mu}_n$ that is constant on the intervals $[0, X_{(1)}]$, $(X_{(n)}, 1]$ and $(X_{(i-1)}, X_{(i)}]$ for all $i = 2, \dots, n-1$.

Now, recall that for every nonincreasing left-continuous function $h : [0,1] \to \mathbb{R}$, the generalized inverse of h is defined as: for every $a \in \mathbb{R}$, $h^{-1}(a)$ is the greatest $t \in [0,1]$ that satisfies $h(t) \geq a$, with the convention that the supremum of an empty set is zero. In the sequel, we consider the generalized inverse $\hat{\mu}_n^{-1}$ of $\hat{\mu}_n$ as an estimator for μ^{-1}.

4.1.2. *Uniform integrability and bias.* Below, we provide bounds on the maximal risk of the isotonic LSE and its inverse, which imply uniform integrability. Although such bounds on the maximal risk over suitable classes of functions are known for most nonparametric function estimators, this is the *first instance* for such a result in the context of isotonic regression. We also establish the order of the bias for both estimators. The proofs are given in Section 7.4.

THEOREM 4.1. *Assume (R4) and that X has a density function f. Let $A_1, \dots, A_5$ be positive numbers and consider $\mathcal{F}_1$, the class of nonincreasing functions μ on $[0,1]$ such that*

$$A_1 \leq \left|\frac{\mu(t) - \mu(x)}{t - x}\right| \leq A_2 \text{ for all } t \neq x \in [0,1], \tag{4.3}$$

$|\mu(t)| < A_5$ for all $t \in [0,1]$, and $A_3 < f(t) < A_4$ for all t. Then, for any $p \geq 1$, there exists $K_p > 0$ that depends only on $p, A_1, \dots, A_5$ and α such that

1. $\limsup_{n\to\infty} \sup_{\mu\in\mathcal{F}_1} n^{p/3}\,\mathbb{E}_\mu\left(|\hat{\mu}_n^{-1}(a) - \mu^{-1}(a)|^p\right) \leq K_p$ *for all fixed $a \in \mathbb{R}$,*
2. $\limsup_{n\to\infty} \sup_{\mu\in\mathcal{F}_1} n^{p/3}\,\mathbb{E}_\mu\left(|\hat{\mu}_n(t) - \mu(t)|^p\right) \leq K_p$ *for all fixed $t \in (0,1)$.*

Note that (4.3) holds if μ has a first derivative that is bounded from both infinity and zero.

REMARK 4.1. *Theorem 4.1 implies that for fixed t, $n^{p/3}\,\mathbb{E}_\mu\left(|\hat{\mu}_n(t) - \mu(t)|^p\right)$ is bounded. This is similar to [15, (11.32) and (11.33)] in the current status model. However, the inequalities in [15] hold for all t, whereas the corresponding inequality for the general regression model above holds only for t in a restricted interval. This is due to a very specific feature of the estimator in the current status model: it has the same range as the estimated function since both of them are distribution functions. In particular, the estimator is consistent at the boundaries in the current status model, whereas it is not in the general regression model. Hence, the strategy of proof in [15] does not extend to our regression model: whereas the proof in [15] is based solely on an exponential inequality for the tail probabilities of the inverse estimator, our proof is based on two tail inequalities, one of them being an extension of Theorem 11.3 in [15] to our setting (Lemma 7.1), and the other one being a sharper inequality for points outside the range of μ (Lemma 7.3).*

We next consider the order of the bias. Tackling the bias requires imposing additional smoothness assumptions on the underlying parameters of the problem. Precisely, we assume for some of our results that v^2 has a bounded second derivative on $[0,1]$, that μ is differentiable with

$$|\mu'(x) - \mu'(y)| \leq C|x-y|^s, \qquad \text{for all } x, y \in [0,1], \tag{4.4}$$

for some $C > 0$ and $s > 0$ (where bounds on s will be specified precisely while stating the actual results); and, instead of (R2), the more restrictive assumption:

(R5) The density f of X is bounded away from zero with a bounded first derivative on $[0,1]$.

THEOREM 4.2. *Assume (R1), (R5), (R3) and (R4). Assume, furthermore, that v^2 has a bounded second derivative on $[0,1]$ and μ satisfies (4.4) for some $C > 0$ and $s > 1/2$. Then,*

$$\mathbb{E}\left(\hat{\mu}_n^{-1}(a) - \mu^{-1}(a)\right) = o(n^{-1/2}) + O(n^{-(2s+3)/9}(\log n)^{25/2})$$

uniformly in $a \in [\mu(1) + Kn^{-1/6}\log n, \mu(0) - Kn^{-1/6}\log n]$.

Now, consider the bias of the direct estimator. For technical reasons, we require a higher degree of smoothness $s = 1$ on μ' than needed for dealing with the inverse function and we obtain a slower rate than for the inverse.

THEOREM 4.3. *Assume (R1), (R5), (R3), (R4), v^2 has a bounded second derivative on $[0,1]$ and (4.4) holds for some $C > 0$ and $s = 1$. For an arbitrary fixed $[c_1, c_2] \subset (0,1)$, we have*

$$\mathbb{E}\left(\hat{\mu}_n(t) - \mu(t)\right) = O(n^{-7/15+\zeta})$$

with an arbitrary $\zeta > 0$, where the big-O term is uniform in $t \in [c_1, c_2]$.

4.1.3. *On the criticality of the smoothness assumption on μ.* Theorems 4.2 and 4.3 show that under appropriate smoothness assumptions, the bias of the isotonic LSE and its inverse converge to 0 at a rate *strictly faster* than $n^{-1/3}$; e.g., in the inverse problem, $n^{1/3}(\mathbb{E}(\hat{\mu}_n^{-1}(a) - \mu^{-1}(a)) = o(c_n^{-1})$ for some c_n going to infinity, whence by Theorem 3.1, we can choose the number of sub-samples $m_n \to \infty$ (in terms of c_n) for constructing the pooled estimator, achieving in the process an acceleration in the convergence rate compared to the global estimator. However, the cube-root convergence rate in the isotonic regression problem does not require smoothness: it is valid even under a Lipschitz assumption on the regression function. It is, therefore, interesting to consider whether the divide and conquer method works under the weaker Lipschitz assumption. This boils down to the question whether the bias of the isotonic estimator (or its inverse) also disappears at a rate faster than $n^{-1/3}$ under Lipschitz continuity. We show in Section 8.13 of the supplement (in the inverse problem setting) that without smoothness, Lipschitz continuity in itself is not sufficient to guarantee a bias that vanishes sufficiently quickly. Indeed, in our example, $n^{1/3}(\mathbb{E}(\hat{\mu}_n^{-1}(a)) - \mu^{-1}(a))$ converges to a non-zero quantity. The pooled estimator therefore accumulates bias and its MSE goes to infinity as m_n increases and divide and conquer *fails* dramatically.

4.1.4. *Sample splitting in the isotonic regression model.* We next study the effect of sample-splitting in the isotonic regression model. We consider N i.i.d. copies $\{(X_i, Y_i)\}_{i=1}^N$ of (X, Y) as above. The parameter of interest is $\theta_0 \equiv \mu(t_0)$ which is estimated by

$$\bar{\theta}_{m_n} = \frac{1}{m_n} \sum_{j=1}^{m_n} \hat{\mu}_{n,j}(t_0),$$

$\hat{\mu}_{n,j}$ being the isotonic LSE computed from the j-th split-sample. Under (a subset of) the assumptions on the parameters of the model made in Theorem 4.3, convergence in law to Chernoff's distribution holds: with $\hat{\mu}$ denoting the global estimator based on all N observations, we have (1.2) with $\kappa := |4v^2(t_0)\mu'(t_0)/f(t_0)|^{1/3}$. To apply Theorem 3.1, we need to show that: (a) $n^{1/3}(\theta_n - \mu(t_0)) = O(n^{-\phi})$ (here $\theta_n = \mathbb{E}[\hat{\mu}_{n,1}(t_0)]$) for some $\phi > 0$, and (b) the uniform integrability of the sequence $\{n^{2/3}(\hat{\mu}_{n,1}(t_0) - \mu(t_0))^2\}_{n \geq 1}$.

Now, (b) is a direct consequence of Theorem 4.1 applied with any $p > 2$. As far as (a) is concerned, by Theorem 4.3, we know that the desired condition in (a) is satisfied for $s = 1$ in (4.4) for any fixed $t_0 \in (0, 1)$, by taking $\phi = (7/15 - 1/3) - \zeta = (2/15 - \zeta)$ where $\zeta > 0$ can be taken to be arbitrarily small. From Remark 3.2, choosing $m_n = n^{2\phi - \delta} = n^{4/15 - 2\zeta - \delta}$ for a small enough $0 < \delta < 2\phi$, we conclude that with $\sigma^2 = \kappa^2 \operatorname{Var}(\mathbb{Z})$, we have

$$N^{(7/15 - \zeta - \delta/2)/(19/15 - 2\zeta - \delta)}(\overline{\theta}_{m_n} - \theta_0) \xrightarrow{d} N(0, \sigma^2)\,. \tag{4.5}$$

4.1.5. *Inverse function estimation at a point.* Consider the same set-up as in Section 4.1.4. We now consider estimation of $\mu^{-1}(a)$ via the inverse isotonic LSE under the assumptions of Theorem 4.2. The behavior of the isotonic estimator $\hat{\mu}^{-1}$ based on the entire data of size N is given by (1.3) where $\tilde{\kappa} := |4v^2(t_0)/\mu'(t_0)^2 f(t_0)|^{1/3}$, with $t_0 = \mu^{-1}(a)$. To apply Theorem 3.1, we need to show that: (a) $n^{1/3}(\theta_n - \mu^{-1}(a)) = O(n^{-\phi})$ (here $\theta_n = \mathbb{E}[\hat{\mu}_{n,1}^{-1}(a)]$) for some $\phi > 0$, and (b) the uniform integrability of the sequence $\{n^{2/3}(\hat{\mu}_{n,1}^{-1}(a) - \mu^{-1}(a))^2\}_{n \geq 1}$.

In this case, (b) is a direct consequence of Theorem 4.1 applied with any $p > 2$. As far as (a) is concerned, by Theorem 4.2, we know that the desired condition in (a) is satisfied for $s > 3/4$

in (4.4) for any fixed a in the interior of the range of μ by taking $\phi = (1/2 - 1/3) = 1/6$. From Remark 3.2, choosing $m_n = n^{2\phi} = n^{1/3}$ (for the inverse function estimation problem we are actually in the situation of conclusion (ii) of Theorem 3.1 with $\tau = 0$), we conclude that:

$$N^{(1/3+1/6)/[2(1/6+1/2)]}(\overline{\theta}_{m_n} - \theta_0) \equiv N^{3/8}(\overline{\theta}_{m_n} - \theta_0) \xrightarrow{d} N(0, \widetilde{\sigma}^2)\,, \tag{4.6}$$

where $\widetilde{\sigma}^2 = \tilde{\kappa}^2 \operatorname{Var}(\mathbb{Z})$. The pooled estimator, therefore, has a convergence rate of $N^{3/8}$.

REMARK 4.2. *The order of the bias obtained in the forward problem (Theorem 4.3) is slower than that obtained in the inverse problem (Theorem 4.2) and comes at the expense of increased smoothness ($s = 1$) compared to Theorem 4.2. This seems to be, at least partly, an artifact of our approach where we start from the characterization of the inverse estimator and derive results for the forward problem from those in the inverse problem through the switching relationship.*

Next, even for the inverse problem, it is not clear at this point whether the order of the bias obtained in Theorem 4.2 is optimal, i.e., the best possible one under the assumed smoothness. It is conceivable that when $s > 3/4$ the exact order of the bias is smaller than the obtained $o(n^{-1/2})$ rate from Theorem 4.2. A smaller bias would allow a faster rate of convergence than $N^{3/8}$ through an appropriate choice of m_n. A complete resolution of the bias problem is outside the scope of this paper. It is, however, worth reiterating that Theorems 4.2 and 4.3 are the first systematic attempts in the literature to quantify the bias of isotonic estimators.

4.2. *The current status model.* The current status model has found extensive applications in epidemiology and biomedicine. One version of this problem is to estimate the distribution function F_T of a failure time $T \geq 0$ on $[0,1]$, based on observing n independent copies of the censored pair $(X, \mathbb{1}_{T\leq X})$. Here, $X \in [0,1]$ is the observation time independent of T, and $\mathbb{1}_{T\leq X}$ stipulates whether or not the failure has occurred before time X. Then,

$$F_T(x) = \mathbb{P}(T \leq x) = \mathbb{E}(\mathbb{1}_{T\leq X} | X = x)$$

for all $x \in [0,1]$. This falls in the general framework of Section 4.1 with $Y = -\mathbb{1}_{T\leq X}$ and $\mu = -F_T$, which is nonincreasing. It follows from [17, Remark page 30] that the NPMLE of F_T is precisely the right-continuous version of $\hat{F}_{Tn} := -\hat{\mu}_n$ where $\hat{\mu}_n$ is the LSE from Section 4.1.1. We give below a separate treatement for $\hat{F}_{Tn}$ in the current status model since $\epsilon := -\mathbb{1}_{T\leq X} + F_T(X)$ does not satisfy the assumption (R4). The following theorem is proved in Section 7.2.

THEOREM 4.4. *Assume that we observe n independent copies of $(X, \mathbb{1}_{T\leq X})$, where $X \in [0,1]$ is independent of $T \geq 0$. Assume that T has a continuous density function f_T that is bounded away from both zero and infinity on $[0,1]$, and that X has a density function f on $[0,1]$ that is bounded away from zero and has a continuous first derivative on $[0,1]$. With $\hat{F}_{Tn}$ as above and the corresponding inverse $\hat{F}_{Tn}^{-1}(a) = \hat{\mu}_n^{-1}(-a)$ we have:*

1. *For any $p \geq 1$, there exists $K_p > 0$ such that for all n, $t \in [0,1]$ and $a \in [0,1]$,*

$$\mathbb{E}\left(|\hat{F}_{Tn}(t) - F_T(t)|^p\right) \leq K_p n^{-p/3} \text{ and } \mathbb{E}\left(|\hat{F}_{Tn}^{-1}(a) - F_T^{-1}(a)|^p\right) \leq K_p n^{-p/3}.$$

2. *If moreover, f_T is Lipschitz continuous, then with arbitrary positive K, c_1, ϕ and $c_2 < 1$,*

$$\mathbb{E}\left(\hat{F}_{Tn}^{-1}(a) - F_T^{-1}(a)\right) = o(n^{-1/2}) \text{ and } \mathbb{E}\left(\hat{F}_{Tn}(t) - F_T(t)\right) = O(n^{-7/15+\phi})$$

uniformly for all $a \in [Kn^{-1/6}\log n, 1 - Kn^{-1/6}\log n]$ and $t \in [c_1, c_2]$.

3. *Now, let $\hat{F}_{TN}$ denote the MLE based on $N = m_n \times n$ observations from the current status model, $\hat{F}_{Tn}^{(j)}$ the MLE from the j'th subsample and $\overline{F}_{m_n}$ the pooled isotonic estimator obtained by averaging the $\hat{F}_{Tn}^{(j)}$s. Under the above assumptions, for all $\zeta, \delta > 0$, sufficiently small, and any $0 < t < 1$, we have*

$$N^{(7/15-\zeta-\delta/2)/(19/15-2\zeta-\delta)}(\overline{F}_{m_n}(t) - F(t)) \xrightarrow{d} N(0, \sigma^2)\,,$$

where $\sigma^2 = \{4\,F_T(t)(1 - F_T(t))f_T(t)/f(t)\}^{2/3}\,\mathrm{Var}(\mathbb{Z})$. Moreover, for any $a \in (0,1)$, with $\overline{\theta}_{m_n}$ the pooled estimator obtained by averaging the $(\hat{F}_{Tn}^{(j)})^{-1}(a)$s,

$$N^{3/8}(\overline{\theta}_{m_n} - F_T^{-1}(a)) \xrightarrow{d} N(0, \tilde{\sigma}^2)\,,$$

where $\tilde{\sigma}^2 = \{4\,a(1-a)/f_T'(t_a)^2 f(t_a)\}^{2/3}\,\mathrm{Var}(\mathbb{Z})$ with $t_a = F_T^{-1}(a)$. On the other hand,

$$N^{1/3}(\hat{F}_{TN}(t) - F(t)) \xrightarrow{d} \{4\,F_T(t)(1 - F_T(t))f_T(t)/f(t)\}^{1/3}\,\mathbb{Z}\,,$$

while

$$N^{1/3}(\hat{F}_{TN}^{-1}(a) - F_T^{-1}(a)) \xrightarrow{d} \{4\,a(1-a)/f_T(t_a)^2 f(t_a)\}^{1/3}\,\mathbb{Z}\,.$$

4.3. *The monotone density and monotone hazard problems.* We further illustrate our results on two widely studied monotone function problems, where the goal is to estimate a function λ on $[0,1]$ under the known constraint that it is nonincreasing.

(a) **Grenander estimator:** Consider i.i.d. data $W_1, \ldots, W_n$ with common nonincreasing density function λ on the interval $[0,1]$. The nonparametric MLE of λ is $\hat{\lambda}_n$, the left-hand slope of the least concave majorant of the empirical distribution function Λ_n corresponding to $W_1, \ldots, W_n$.

(b) **Monotone hazard under right censoring:** Consider i.i.d. data $\{W_i := (X_i, \delta_i) : i = 1, \ldots, n\}$ from the random censorship model: $X_i = \min(T_i, C_i)$ and $\delta_i = \mathbb{1}_{\{T_i \le C_i\}}$. The failure times T_i are assumed to be nonnegative with density f and to be independent of the i.i.d. censoring times C_i that have a distribution function G. The failure rate $\lambda = f/(1-F)$, where F is the distibution function corresponding to f, is assumed to be nonincreasing. We will consider the Huang-Wellner estimator $\hat{\lambda}_n$ on $[0,1]$, defined as the left-hand slope of the least concave majorant of the restriction of the Nelson-Aalen estimator Λ_n to $[0,1]$ [18]. Recall that if $t_1 < \cdots < t_K$ denote the ordered distinct uncensored failure times, and n_k the number of $i \in \{1, \ldots, n\}$ with $X_i \ge t_k$, then Λ_n is constant on $[t_k, t_{k+1})$ with $\Lambda_n(t_k) = \sum_{i \le k} n_i^{-1}$ for all k, $\Lambda_n(t) = 0$ for all $t < t_1$, and $\Lambda_n(t) = \Lambda_n(t_K)$ for all $t \ge t_K$.

Given some $t_0 \in (0,1)$ the parameter of interest is $\theta_0 = \lambda(t_0)$.

THEOREM 4.5. *Assume that we observe n independent copies of W in either the framework (a) or (b), with the corresponding estimator $\hat{\lambda}_n$ of λ. Assume that λ is differentiable and decreasing on $[0,1]$ with $\inf_t \lambda(t) > 0$ and*

$$A_1 \le \left|\frac{\lambda(t) - \lambda(x)}{t - x}\right| \le A_2 \text{ for all } t \ne x \in [0,1], \tag{4.7}$$

In addition, under (b), assume that $F(1) < 1$, $\lim_{t \uparrow 1} G(t) < 1$, and G has a bounded continuous derivative on $(0,1)$. We then have, in both settings (a) and (b):

1. *For any $p \geq 1$, there exists $K_p > 0$ such that for all n, $t \in [n^{-1/3}, 1 - n^{-1/3}]$ and $a \in \mathbb{R}$,*

$$\mathbb{E}\left(|\hat{\lambda}_n(t) - \lambda(t)|^p\right) \leq K_p n^{-p/3} \text{ and } \mathbb{E}\left(|\hat{\lambda}_n^{-1}(a) - \lambda^{-1}(a)|^p\right) \leq K_p n^{-p/3}.$$

2. *If moreover, λ has a first derivative that satisfies $|\lambda'(t) - \lambda'(x)| \leq A|t - x|$ for all $t, x \in [0,1]$ and some $A > 0$, then with $K > 0$, $c_1 > 0$, $c_2 < 1$, and $\phi > 0$ arbitrary constants,*

$$\mathbb{E}\left(\hat{\lambda}_n^{-1}(a) - \lambda^{-1}(a)\right) = o(n^{-1/2}) \text{ and } \mathbb{E}\left(\hat{\lambda}_n(t) - \lambda(t)\right) = O(n^{-7/15+\phi})$$

uniformly for all $a \in [Kn^{-1/6}\log n, 1 - Kn^{-1/6}\log n]$ and $t \in [c_1, c_2]$.

3. *Now, let $\hat{\lambda}_N$ denote the MLE based on $N = m_n \times n$ observations, $\hat{\lambda}_n^{(j)}$ the MLE from the j'th subsample and $\overline{\lambda}_{m_n}$ the pooled isotonic estimator obtained by averaging the $\hat{\lambda}_n^{(j)}$s. Under the above assumption, for all $\zeta, \delta > 0$ sufficiently small, and any $0 < t < 1$, we have*

$$N^{(7/15-\zeta-\delta/2)/(19/15-2\zeta-\delta)}(\overline{\lambda}_{m_n}(t) - \lambda(t)) \xrightarrow{d} N(0, \kappa^2 \mathrm{Var}(\mathbb{Z})),$$

where $\kappa = |4\lambda(t)\lambda^{'}(t)|^{1/3}$ under (a) and $\kappa = |4\lambda(t)\lambda^{'}(t)/\overline{H}(t)|^{1/3}$ under (b), where $\overline{H}(t) = (1 - F(t))(1 - G(t))$. Moreover, for any $a \in (0,1)$, with $\overline{\theta}_{m_n}$ being the pooled estimator obtained by averaging the $(\hat{\lambda}_n^{(j)})^{-1}(a)$s,

$$N^{3/8}(\overline{\theta}_{m_n} - \lambda^{-1}(a)) \xrightarrow{d} N(0, \tilde{\kappa}^2 \mathrm{Var}(\mathbb{Z})),$$

with $\tilde{\kappa} = |4a/[\lambda^{'}(\lambda^{-1}(a))]^2|^{1/3}$ under (a) and $\tilde{\kappa} = |4a/[\lambda^{'}(\lambda^{-1}(a))]^2\,\overline{H}(\lambda^{-1}(a))|^{1/3}$ under (b). On the other hand:

$$N^{1/3}(\hat{\lambda}_N(t_0) - \lambda(t_0)) \xrightarrow{d} \kappa\mathbb{Z},$$

while

$$N^{1/3}(\hat{\lambda}_N^{-1}(a) - \lambda^{-1}(a)) \xrightarrow{d} \tilde{\kappa}\mathbb{Z}.$$

To save space, and given that the proof techniques are similar to those of the models considered earlier, an outline of the proof of the above theorem is relegated to Section 8.14 of the supplement.

5. Sample-splitting and the super-efficiency phenomenon.

The variance reduction accomplished by sample-splitting (see (2.2)) for estimating a fixed monotone function, or its inverse, at a given point comes at a price. We show in this section, in the context of the inverse isotonic regression problem, that though a larger number of splits (m) brings about greater reduction in the variance for a fixed function, the performance of the pooled estimator in a *uniform sense*, over an appropriately large class of functions, deteriorates in comparison to the global estimator as m increases. This can be viewed as a *super-efficiency* phenomenon: a trade-off between point wise performance and performance in a uniform sense.

5.1. *Super-efficiency of the pooled estimator.*

Fix a nonincreasing function μ_0 on $[0,1]$ that is continuously differentiable on $[0,1]$ with $0 < c < |\mu_0^{'}(t)| < d < \infty$ for all $t \in [0,1]$. Let $x_0 \in (0,1)$. Define a neighborhood $\mathcal{M}_0$ of μ_0 as the class of all continuous nonincreasing functions μ on $[0,1]$ that are continuously differentiable on $[0,1]$, that coincide with μ_0 outside of $(x_0 - \epsilon_0, x_0 + \epsilon_0)$ for some (small) $\epsilon_0 > 0$, and such that $0 < c < |\mu'(t)| < d < \infty$ for all $t \in [0,1]$. Now, consider

N i.i.d. observations $\{(X_i, Y_i)\}_{i=1}^N$ from (X, Y) as in Section 4.1 where $X \sim \text{Uniform}(0,1)$ is independent of $\epsilon \sim N(0, v^2)$. We know that the isotonic estimate $\hat{\theta}_N$ of $\theta_0 := \mu_0^{-1}(a)$ satisfies

$$N^{1/3}\,(\hat{\theta}_N - \theta_0) \xrightarrow{d} G, \tag{5.1}$$

as $N \to \infty$, where $G =_d \tilde{\kappa}\,\mathbb{Z}$, $\mathbb{Z}$ being the Chernoff random variable, and $\tilde{\kappa} > 0$ is a constant. If we split N as $m \times n$, where m is a fixed integer, then as $N \to \infty$, Lemma 2.1 tells us that $N^{1/3}(\overline{\theta}_m - \theta_0)$ converges in distribution to $m^{-1/6}H$, where $\overline{\theta}_m$ is the pooled estimator and H has the same variance as G. By Theorem 4.1 we have uniform integrability under μ_0, whence

$$\mathbb{E}_{\mu_0}\left[N^{2/3}(\hat{\theta}_N - \theta_0)^2\right] \to \text{Var}(G) \quad \text{and} \quad \mathbb{E}_{\mu_0}\left[N^{2/3}(\overline{\theta}_m - \theta_0)^2\right] \to m^{-1/3}\,\text{Var}(G), \tag{5.2}$$

as $N \to \infty$. Hence, the pooled estimator *outperforms the inverse* isotonic regression estimator.

We now focus on comparing the performance of the two estimators over the class $\mathcal{M}_0$. In this regard we have the following theorem, proved in Section 7.3.

THEOREM 5.1. *Let*

$$E := \limsup_{N\to\infty} \sup_{\mu\in\mathcal{M}_0} \mathbb{E}_\mu\left[N^{2/3}(\hat{\theta}_N - \theta_0)^2\right] \quad \textit{and} \quad E_m := \liminf_{N\to\infty} \sup_{\mu\in\mathcal{M}_0} \mathbb{E}_\mu\left[N^{2/3}(\overline{\theta}_m - \theta_0)^2\right] \tag{5.3}$$

where the subscript m indicates that the maximal risk of the m-fold pooled estimator (m fixed) is being considered. Then $E < \infty$ while $E_m \geq m^{2/3}\,c_0$, for some $c_0 > 0$. When $m = m_n$ diverges to infinity,

$$\liminf_{N\to\infty} \sup_{\mu\in\mathcal{M}_0} \mathbb{E}_\mu\left[N^{2/3}(\overline{\theta}_{m_n} - \theta_0)^2\right] = \infty\,.$$

Therefore, from Theorem 5.1 it follows that the asymptotic maximal risk of the pooled estimator diverges to ∞ (at least) at rate $m^{2/3}$. Thus, the better off we are in a pointwise sense with the pooled estimator, the worse off we are in the uniform sense over the class of functions $\mathcal{M}_0$.

REMARK 5.1. *As an inspection of the proof of this theorem reveals, the super-efficiency phenomenon with this dichotomy of pointwise and uniform risk is really an outcome of a bias-related problem. The maximal squared bias of the appropriately normalized pooled estimator over the class of functions $\mathcal{M}_0$ considered in the above theorem diverges to ∞ owing to the fact that the maximal squared bias (over the class $\mathcal{M}_0$) of the subsample level isotonic estimates fails to go to 0. Essentially, the class $\mathcal{M}_0$ is so large that the Hölder condition* (4.4) *is not satisfied uniformly over $\mathcal{M}_0$.*

Table 1 gives the ratios of the (estimated) mean squared errors $\mathbb{E}\left[(\hat{\mu}_N^{-1}(a) - \theta_0)^2\right]/\mathbb{E}\left[(\overline{\theta}_m - \theta_0)^2\right]$ comparing the performance of the pooled estimator $\overline{\theta}_m$ with the global estimator $\hat{\mu}_N^{-1}(a)$ as n and m change for two different models, which are described in the caption to the table. For the first model (left table) we fix $\mu(x) = x$ and let $N \to \infty$ and find that the pooled estimator has superior performance to the global estimator as m (and n) grows. The ratio of the mean squared errors is generally close to $m^{1/3}$, as per (5.2). The second model considered (right table) illustrates the phenomenon described in Theorem 5.1. We lower bound the supremum risk over $\mathcal{M}_0$ by considering a sequence of alternatives in $\mathcal{M}_0$ (obtained from local perturbations to

(n, m)	5	10	15	30	45	60	90	5	10	15	30	45	60	90
50	1.67	1.71	1.90	1.66	1.57	1.65	1.17	1.47	1.21	0.94	0.70	0.55	0.54	0.39
100	1.31	1.76	2.21	2.29	2.16	2.46	2.33	1.04	0.97	0.90	0.59	0.47	0.40	0.31
200	1.75	2.06	2.42	2.81	2.58	3.16	3.39	1.03	0.94	0.76	0.68	0.42	0.38	0.29
500	1.70	2.13	2.12	2.80	3.16	3.59	4.11	1.01	0.90	0.69	0.54	0.44	0.34	0.24
1000	1.46	2.04	2.46	2.88	3.60	3.51	4.31	1.16	0.88	0.66	0.52	0.36	0.34	0.24
3000	1.63	2.12	2.33	3.11	4.15	3.84	3.69	1.09	0.87	0.75	0.43	0.40	0.31	0.21
10000	1.75	2.11	2.70	2.86	3.31	5.08	5.18	0.94	0.79	0.80	0.43	0.33	0.31	0.23

TABLE 1

Ratios of the (estimated) mean squared errors $\frac{\mathbb{E}[(\hat{\mu}_N^{-1}(a)-\theta_0)^2]}{\mathbb{E}[(\bar{\theta}_m-\theta_0)^2]}$ *comparing the performance of the pooled estimator* $\bar{\theta}_m$ *with the global estimator* $\hat{\mu}_N^{-1}$ *as* n *and* m *change for the model:* $Y = \mu(X) + \epsilon$, $X \sim Unif(0, 1)$, $\epsilon \sim N(0, 0.2^2)$, *and* $a = 0.5$, *with (i)* $\mu(x) = x$, *and (ii)* $\mu(x) = \mu_n(x) = x + n^{-1/3}B(n^{1/3}(x - x_0))$ *with* $B(u) = 2^{-1}(1 - (|u| - 1)^2)^2 \mathbb{1}_{\{|u|\leq 2\}}$. *For both (i) and (ii),* $\theta_0 \equiv \mu^{-1}(a) = 0.5$.

$\mu(x) = x$ around $x_0 = 0.5$) for which the ratio of the mean squared errors falls dramatically below 1, suggesting that in such a scenario it is better to use the global estimator $\hat{\mu}_N^{-1}(a)$.

The super-efficiency phenomenon noted in connection with the pooled estimator in the monotone regression model is also seen with sample-splitting with smoothing based procedures, e.g., kernel based estimation, if the bandwidth used in the divide and conquer method is not appropriately adjusted. We describe the phenomenon in a density estimation setting, since this is the easiest to deal with, in Section 8.15 of the supplementary document. Indeed, several authors have criticized such super-efficiency phenomena in nonparametric function estimation; see e.g., [6], [27, Section 1.2.4]. Indeed, it is shown in the second reference that (under the usual twice differentiability assumptions) there exist infinitely many bandwidths that, under any fixed density, produce kernel estimates with asymptotically strictly smaller MSE than the *Epanechnikov oracle* and argued therein that the criterion of assessment of an estimator should therefore be quantified in terms of its maximal risk over an entire class of densities.

While this is certainly a reasonable perspective, we believe that there is also some merit in studying the pointwise behavior of estimators such as in (5.1) (as opposed to a uniform measure such as (5.3)). For construction of CIs statisticians usually rely on such pointwise asymptotic results as it is often quite difficult to obtain useful practical procedures that have justification in a uniform sense. Moreover, in the regime of massive datasets, sample-splitting can provide practical gains over the global estimator which might be impossible to compute.

6. Conclusion.

We have established rigorous results on the behavior of the pooled (by averaging) estimator using sample-splitting in a variety of nonparametric monotone function estimation problems and demonstrated both its pros and cons. The dichotomy between pointwise risk and maximal risk demonstrated in this paper is expected to arise more broadly in many of the other cube-root M-estimation problems mentioned in the Introduction and developed in [19], since the inverse monotone regression problem treated in this paper is as an M-estimation problem of the type considered in [19]. A generic treatment of this class of problems should provide an interesting avenue for future research but is outside the scope of this paper. A more general (and harder) question worth considering is a broad characterization of non-standard problems (not necessarily with cube-root convergence rates) where sample-splitting followed by averaging improves the point-wise risk but produces out-of-control uniform risk bounds, and also how one can circumvent this dichotomy by the use of other clever divide-and-conquer algorithms.

7. Proofs of the main results.

7.1. *Proof of Theorem 3.1.* Since $\{\xi_{n,1}^2\}_{n\geq 1}$ is uniformly integrable and $\xi_{n,1} \xrightarrow{d} G$, $\sigma_n^2 := \mathrm{Var}(\xi_{n,1}) \to \sigma^2$ as $n \to \infty$. Set

$$Z_n := \sum_{j=1}^{m_n} (\xi_{n,j} - b_n)$$

and let $B_n^2 := \mathrm{Var}(Z_n) = m_n \sigma_n^2$. Now, with $\bar{\xi}_n = m_n^{-1} \sum_{j=1}^{m_n} \xi_{n,j}$ we have

$$\frac{Z_n}{B_n} = \frac{\sum_{j=1}^{m_n}(\xi_{n,j} - b_n)}{\sqrt{m_n}\sigma_n} = \frac{\sqrt{m_n} r_n(\bar{\theta}_{m_n} - \theta_0)}{\sigma_n} - \frac{\sqrt{m_n}\, b_n}{\sigma_n} \equiv I_n - II_n.$$

To show that $Z_n/B_n \xrightarrow{d} N(0,1)$, we just need to verify the Lindeberg condition: for every $\epsilon > 0$,

$$\frac{1}{\sigma_n^2}\,\mathbb{E}[(\xi_{n,1} - b_n)^2 \mathbf{1}\{|\xi_{n,1} - b_n| > \epsilon\sqrt{m_n}\,\sigma_n\}] \to 0.$$

Since σ_n^2 converges to $\sigma^2 > 0$ and $m_n \to \infty$, the above condition is implied by the uniform integrability of $\{(\xi_{n,1} - b_n)^2\}_{n\geq 1}$ which is guaranteed by the uniform integrability of $\{\xi_{n,1}^2\}$ (since the sequence b_n goes to 0 and is therefore bounded). Hence, $Z_n/B_n \xrightarrow{d} N(0,1)$.

Now assume that m_n is as in (i). Then, $II_n \to 0$, which implies that I_n converges to a standard Gaussian law, whence (i) follows. Next, if m_n is as in (ii), $II_n \to \tau/\sigma$, and (ii) follows. □

7.2. *Proof of Theorem 4.4.* Let $\mu = -F_T$ and for all $i = 1, \ldots, n$, let $Y_i = -\mathbb{1}_{T_i \leq X_i}$ and $\epsilon_i = Y_i - \mu(X_i) \in [-1, 1]$. Moreover, define $v^2(x) := \mathbb{E}(\epsilon_i^2 | X_i = x)$ for all $x \in [0,1]$. We then have

$$v^2(x) = \mathrm{Var}(\mathbb{1}_{T \leq x}) = F_T(x)(1 - F_T(x)).$$

Note that $F_T^{-1}(a) = \mu^{-1}(-a)$. Under the assumptions of Theorem 4.4, (R1) and (R5) hold true. The assumption (R3) holds since T has a density function that is bounded away from zero. However, (R4) does not hold so Theorems 4.1, 4.2 and 4.3 cannot be directly applied to obtain the results in the current status model. Nevertheless, we will follow the same line of proof in the current status model as in the general regression model. Note that in the current setting, the variance function v^2 may not have a bounded second derivative but instead, it has a Lispschitz first derivative with is in fact enough for our purposes. Hence, the first step is to obtain analoguous of the preliminary lemmas of Section 7.4.1 for the current status model. As a consequence of Theorem 11.3 in [15], the inequality (7.8) still holds for all $a \in [0,1]$ and $x > 0$. Because $\hat{\mu}_n^{-1}(a) = \mu^{-1}(a)$ for all $a \notin [0,1]$ in the current status case, the inequality also holds for all $a \in \mathbb{R}$. Because thanks to (7.4), $\hat{U}_n = F_n(\hat{\mu}_n^{-1}) + O(n^{-1})$, combining this with Corollary 1 in [22] implies that (7.9) also holds for all $a \in \mathbb{R}$ and $x > 0$. Now, $\hat{\mu}_n^{-1}(a)$ is equal ot 0 for all $a > \lambda(0)$ and to 1 for all $a < \lambda(1)$, so (7.10) holds for all $a > \lambda(0)$ and $x > n^{-1}$ (since the probability on the left-hand side is zero), whereas (7.11) holds for all $a > \lambda(1)$ and $x > n^{-1}$. This means that one can still apply Lemmas 7.1, 7.2 and 7.3 in the current status model.

The second assertion of Theorem 4.4 follows from Theorem 11.3 in [15]. The first one follows from (11.32) and (11.33) of that book. The conclusions in 2 (on the orders of the bias of $\hat{F}_{Tn}$ and $\hat{F}_{Tn}^{-1}$) follow by the same arguments as for the proof of Theorems 4.2 and 4.3 using that the preparatory lemmas of Section 7.4.1 still apply and $|\epsilon_i| \leq 1$, and the conclusions in 3 follow exactly in the same fashion as for the general regression model considered above. □

7.3. *Proof of Theorem 5.1.* Here, we assume that μ_0 is *nondecreasing* — this is convenient as we borrow several results from other papers stated in the context when μ_0 is nondecreasing. The neighborhood $\mathcal{M}_0$ in the statement of the theorem needs to be similarly modified. (Of course, appropriate changes will lead to the proof of the case when μ_0 is nonincreasing.)

By Theorem 4.1 (adapted to nondecreasing functions), with $p = 2$ and noting that $\mathcal{M}_0$ is a subset of an appropriate $\mathcal{F}_1$ we conclude that $E < \infty$. Letting

$$V_1 := \limsup_{N\to\infty} \sup_{\mu\in\mathcal{M}_0} \mathrm{Var}_\mu[N^{1/3}(\hat{\theta}_N - \mu^{-1}(a))],$$

and

$$V_2 := \limsup_{N\to\infty} \sup_{\mu\in\mathcal{M}_0} N^{2/3}[\mathbb{E}_\mu\, \hat{\theta}_N - \mu^{-1}(a)]^2\,,$$

we have $V_1 \vee V_2 \le E \le V_1 + V_2 < \infty$. Recall that as $\overline{\theta}_m$ is the average of the m i.i.d. random variables $\hat{\mu}_{n,j}^{-1}(a)$, $j = 1, \ldots, m$, $\mathbb{E}_\mu(\overline{\theta}_m) = \mathbb{E}_\mu(\hat{\mu}_{n,1}^{-1}(a))$. Now, consider

$$\begin{aligned} V_{2,m} \quad &:= \quad \liminf_{N} \sup_{\mu\in\mathcal{M}_0} N^{2/3}[\mathbb{E}_\mu\, \overline{\theta}_m - \mu^{-1}(a)]^2 \\ &= \quad m^{2/3} \liminf_{n\to\infty} \sup_{\mu\in\mathcal{M}_0} n^{2/3}[\mathbb{E}_\mu\, \hat{\mu}_{n,1}^{-1}(a) - \mu^{-1}(a)]^2 \quad =: \quad m^{2/3}\, \widetilde{V}_2. \end{aligned}$$

Note that, $E_m \ge V_{2,m} = m^{2/3}\, \widetilde{V}_2$. We will show below that $\widetilde{V}_2 > 0$; thus c_0 in the statement of the theorem can be chosen to be $\widetilde{V}_2$. To this end, consider the monotone regression model under a sequence of *local alternatives* μ_n which eventually lie in $\mathcal{M}_0$. Let $Y = \mu_n(X) + \epsilon$ where everything is as before but μ_0 changes to μ_n which is defined as

$$\mu_n(x) = \mu_0(x) + n^{-1/3} B\left(n^{1/3}(x - \theta_0)\right)$$

and B is *a non-zero function* continuously differentiable on $\mathbb{R}$, vanishing outside $(-1, 1)$, such that μ_n is monotone for each n and lies eventually in the class $\mathcal{M}_0$[3]. Note that μ_n and μ_0 can differ on $(\theta_0 - n^{-1/3}, \theta_0 + n^{-1/3})$ only, and that $\mu_n'(x) = \mu_0'(x) + B'(n^{1/3}(x - \theta_0))$ for $x \in [\theta_0 - n^{-1/3}, \theta_0 + n^{-1/3}]$ and $\mu_n'(x) = \mu_0'(x)$ otherwise. It is clear that this can be arranged for infinitely many B's.

The above sequence of local alternatives was considered in [1] in a more general setting, namely that of monotone response models, where (in a somewhat unfortunate collision of notation) X denotes *response* and Z the covariate. We invoke the results of that paper *using the* (Y, X) *notation of this paper and ask the reader to bear this in mind.* Using our current notation for the problem in [1], X follows density $p_X(x) = \mathbb{I}_{(0,1)}(x)$ and $Y \mid X = x \sim p(y, \psi(x))$, ψ being a monotone function and $p(y, \theta)$ a regular parametric model. The monotone regression model with homoscedastic normal errors under current consideration is a special case of this setting with $p(y, \theta)$ being the $N(\theta, v^2)$ density, the ψ_n's in that paper defining the local alternatives are the monotone functions μ_n, $\psi_0 = \mu_0$, $c = 1$ and $A_n(x) = B(n^{1/3}(x - x_0))$ for all n. Invoking Theorems 1 and 2 of [1] with the appropriate changes, we conclude that under μ_n,

$$X_n(h) := n^{1/3}(\hat{\mu}_n(\theta_0 + hn^{-1/3}) - \mu_0(\theta_0)) \xrightarrow{d} g_{c,d,\mathcal{D}}(h),$$

[3] There is nothing special about $(-1, 1)$ as far as constructing the B is concerned. Any $(-c, c)$, for $c > 0$ can be made to work.

where $c = v, d = \mu_0'(x_0)/2$, $\mathcal{D}$ is a shift function given by[4]:

$$\mathcal{D}(t) = \left(\int_0^{t\wedge 1} B(u)du\right) \mathbb{1}_{(0,\infty)}(t) - \left(\int_{t\vee -1}^{0} B(u)du\right) \mathbb{1}_{(-\infty,0)}(t),$$

and $g_{c,d,\mathcal{D}}$ is the right-derivative process of the greatest convex minorant (GCM) of $X_{c,d,\mathcal{D}}(t) := cW(t) + dt^2 + \mathcal{D}(t)$ with W being a two-sided Brownian motion. Now, by essentially the same calculation as on Page 422 of [5],

$$P(n^{1/3}[\hat{\mu}_n^{-1}(a+\lambda n^{-1/3})-\mu_0^{-1}(a)] \le x) = P(n^{1/3}(\hat{\mu}_n(\theta_0+xn^{-1/3})-\mu_0(\theta_0)) \ge \lambda) \to P(g_{c,d,\mathcal{D}}(x) \ge \lambda)\,.$$

Setting $\lambda = 0$, we get:

$$P(n^{1/3}[\hat{\mu}_n^{-1}(a) - \mu_0^{-1}(a)] \le x) = P(n^{1/3}(\hat{\mu}_n(\theta_0 + xn^{-1/3}) - \mu_0(\theta_0)) \ge 0) \to P(g_{c,d,\mathcal{D}}(x) \ge 0)\,.$$

Next, by the switching relationship[5],

$$P(g_{c,d,\mathcal{D}}(x) \ge 0) = P(\arg\min_h X_{c,d,\mathcal{D}}(h) \le x)\,,$$

and it follows that:

$$n^{1/3}(\hat{\mu}_n^{-1}(a) - \mu_0^{-1}(a)) \xrightarrow{d} \arg\min_h\ X_{c,d,\mathcal{D}}(h)\,.$$

Choosing B such that $B(0) = 0$, we note that $\mu_n^{-1}(a) = \mu_0^{-1}(a) = \theta_0$, and therefore, under the sequence of local alternatives μ_n,

$$n^{1/3}(\hat{\mu}_n^{-1}(a) - \mu_n^{-1}(a)) \xrightarrow{d} \arg\min_h X_{c,d,\mathcal{D}}(h). \tag{7.1}$$

Since the μ_n's eventually fall within the class $\mathcal{M}_0$, by Theorem 4.1 (adapted to nondecreasing functions), we conclude that:

$$\limsup_{n\to\infty} n^{2/3}\, \mathbb{E}_{\mu_n}\left(|\hat{\mu}_n^{-1}(a) - \mu_n^{-1}(a)|^2\right) \le K_2.$$

Thus the sequence $\{n^{1/3}(\hat{\mu}_n^{-1}(a) - \mu_n^{-1}(a))\}_{n\ge 1}$ is uniformly integrable under the sequence (of probability distributions corresponding to) $\{\mu_n\}_{n\ge 1}$ and in conjunction with (7.1) it follows that

$$\lim_{n\to\infty}\ n^{1/3}[\mathbb{E}_{\mu_n}(\hat{\mu}_n^{-1}(a) - \mu_n^{-1}(a))] = \mathbb{E}(\arg\min_h X_{c,d,\mathcal{D}}(h))\,.$$

[**Claim C**] (proved in Section 8.16 of the Supplement): For any non-negative function B that satisfies the conditions imposed above, and is additionally symmetric about 0,

$$\mathbb{E}(\arg\min_h X_{c,d,\mathcal{D}}(h)) \neq 0\,.$$

It follows that for any such B,

$$[\mathbb{E}(\arg\min_h X_{c,d,\mathcal{D}}(h))]^2 \le \widetilde{V}_2,$$

[4] There is a typo in the drift term as stated on page 514 of [1]: there should be a negative sign before the integral that defines $\mathcal{D}(h)$ for $h < 0$ on page 514.

[5] For the details, see Section 8.16 of the Supplementary Material.

and hence $\widetilde{V}_2 > 0$. This delivers the assertions of the theorem for fixed m.

When $m = m_n \to \infty$, note that

$$\begin{aligned}\liminf_{N\to\infty} \sup_{\mu\in\mathcal{M}_0} \mathbb{E}_\mu \left[N^{2/3}(\overline{\theta}_{m_n} - \mu^{-1}(a))^2\right] &\geq \liminf_{N\to\infty} \sup_{\mu\in\mathcal{M}_0} N^{2/3}[\mathbb{E}_\mu \overline{\theta}_{m_n} - \mu^{-1}(a)]^2 \\ &\geq \liminf_{n\to\infty} m_n^{2/3} \sup_{\mu\in\mathcal{M}_0} n^{2/3}[\mathbb{E}_\mu \overline{\theta}_{m_n} - \mu^{-1}(a)]^2 \\ &= \liminf_{n\to\infty} m_n^{2/3} \sup_{\mu\in\mathcal{M}_0} n^{2/3}[\mathbb{E}_\mu \hat{\mu}_{n,1}^{-1}(a) - \mu^{-1}(a)]^2 .\end{aligned}$$

By our derivations above,

$$\sup_{\mu\in\mathcal{M}_0} n^{2/3}[\mathbb{E}_\mu \hat{\mu}_{n,1}^{-1}(a) - \mu^{-1}(a)]^2 \geq \frac{1}{2}[\mathbb{E}(\arg\min_h X_{c,d,\mathcal{D}}(h))]^2 > 0$$

for all sufficiently large n. Hence, the liminf of the maximal normalized risk of $\overline{\mu}_N$ is infinite. □

7.4. *Some Selected Proofs for Section 4.1.2.* From (4.2) we have

$$\hat{\mu}_n(X_{(i)}) = \hat{\lambda}_n(i/n) = \hat{\lambda}_n \circ F_n(X_{(i)}), \qquad i = 1, \ldots, n, \tag{7.2}$$

where F_n is the empirical distribution function of $X_1, \ldots, X_n$. We will first study $\hat{\lambda}_n$ and then go back to $\hat{\mu}_n$ thanks to (7.2). Note that $\hat{\lambda}_n(i/n) = \hat{\mu}_n \circ F_n^{-1}(i/n)$ for all $i \in \{1, \ldots, n\}$, where $F_n^{-1}(a)$ is the smallest $t \in [0,1]$ that satisfies $F_n(t) \geq a$, for all $a \in \mathbb{R}$. Both functions $\hat{\lambda}_n$ and $\hat{\mu}_n \circ F_n^{-1}$ are piecewise constant, so $\hat{\lambda}_n = \hat{\mu}_n \circ F_n^{-1}$ on $[0,1]$ and $\hat{\lambda}_n$ estimates

$$\lambda := \mu \circ F^{-1}. \tag{7.3}$$

Let μ^{-1} and g be the respective generalized inverses of μ and λ, which extend the usual inverses to the whole real line in such a way that they remain constant on $(-\infty, 0]$ and on $[1, \infty)$. Letting $\hat{\mu}_n^{-1}$ and $\hat{U}_n$ be the respective generalized inverses of $\hat{\mu}_n$ and $\hat{\lambda}_n$, it follows from (7.2) that

$$\hat{\mu}_n^{-1} = F_n^{-1} \circ \hat{U}_n, \tag{7.4}$$

and it can be shown that

$$\hat{U}_n(a) = \operatorname*{argmax}_{u\in[0,1]}\{\Lambda_n(u) - au\}, \qquad \text{for all } a \in \mathbb{R} \tag{7.5}$$

where argmax denotes the greatest location of maximum (which is achieved on the set $\{i/n,\ i = 0, \ldots, n\}$ since Λ_n is piecewise-linear). Part of the proofs below consist in first establish a result for $\hat{U}_n$ using the above characterization, and then go from $\hat{U}_n$ to $\hat{\mu}_n^{-1}$ using (7.4). To this end, we will use a precise bound for the uniform distance between F^{-1} and F_n^{-1}, as well as a strong approximation of the empirical quantile function, see Section 8.1 in the supplementary material.

We will repeatedly use the fact that because $g' = 1/\lambda' \circ g$ on $(\lambda(1), \lambda(0))$ where $\lambda' = \mu' \circ F^{-1}/f \circ F^{-1}$ is bounded away from zero under (R1) and (R2), for all $u, v \in \mathbb{R}$ we have

$$|g(u) - g(v)| \leq \frac{1}{\inf_{t\in[0,1]} |\lambda'(t)|}|u - v|. \tag{7.6}$$

Furthermore, we recall that Fubini's theorem implies that for all random variables Z and $r \geq 1$,

$$\mathbb{E}|Z|^r = \int_0^\infty \mathbb{P}(|Z|^r > x)dx = \int_0^\infty \mathbb{P}(|Z| > t)rt^{r-1}dt. \tag{7.7}$$

We denote by $\mathbb{P}^X$ and $\mathbb{E}^X$ the conditional probability and expectation given $(X_1, \ldots, X_n)$.

7.4.1. *Preliminaries.* In this section, we provide exponential bounds, which are proved in the supplementary material, for the tail probabilities of $\hat{\mu}_n^{-1}$ and $\hat{U}_n$. We begin with a generalization to our setting of Theorem 11.3 in [15]. Also, the lemma is a stronger version of inequality (11) in [11] where an assumption (A5) was postulated instead of the stronger assumption (R4).

Lemma 7.1. *Assume (R4), X has a density function f, μ is nonincreasing and there exist positive numbers $A_1, \ldots, A_4$ such that (4.3) holds and $A_3 < f(t) < A_4$ for all $t \in [0,1]$. Then, there exist positive numbers K_1 and K_2 that depend only on $A_1, \ldots, A_4, \alpha$, where α is taken from (R4), such that for all n, $a \in \mathbb{R}$ and $x > 0$, we have*

$$\mathbb{P}\left(|\hat{\mu}_n^{-1}(a) - \mu^{-1}(a)| > x\right) \leq K_1 \exp(-K_2 n x^3). \tag{7.8}$$

To prove Lemma 7.1, we first prove a similar bound for $\hat{U}_n$. The exponential bound for $\hat{U}_n$ is given in the following lemma. It will be used also in the proof of Theorem 4.1.

Lemma 7.2. *Under the assumptions of Lemma 7.1, there exist positive numbers K_1 and K_2 that depend only on $A_1, \ldots, A_4$ and α such that for all n, $a \in \mathbb{R}$ and $x > 0$, we have*

$$\mathbb{P}\Big(|\hat{U}_n(a) - g(a)| > x\Big) \leq K_1 \exp(-K_2 n x^3). \tag{7.9}$$

To prove Theorem 4.1, we also need a sharper inequality for the cases when $a \notin [\lambda(1), \lambda(0)]$.

Lemma 7.3. *Assume (R4), X has a density function f, and μ is nonincreasing. Then, there exist positive numbers K_1 and K_2 that depend only α, which is taken from (R4), such that*

$$\mathbb{P}^X\Big(\hat{U}_n(a) \geq x\Big) \leq K_1 \exp(-K_2 (a - \lambda(0))^2 n x) \tag{7.10}$$

for all n, $a > \lambda(0)$ and $x > n^{-1}$, and

$$\mathbb{P}^X\Big(1 - \hat{U}_n(a) \geq x\Big) \leq K_1 \exp(-K_2 (a - \lambda(1))^2 n x) \tag{7.11}$$

for all n, $a < \lambda(1)$ and $x > n^{-1}$.

7.4.2. *Proof of Theorem 4.1.* Integrating the inequality in Lemma 7.1 according to (7.7) proves the first assertion. To prove the second one, we first prove a similar result for $\hat{\lambda}_n$.

Lemma 7.4. *Under the assumptions of Lemma 7.1, for all $p > 0$ and $A > 0$, there exist positive K_1, K_2 that depend only on $A_1, \ldots, A_4, \alpha, p$ and A such that for all n and $t \in [n^{-1/3}A, 1-n^{-1/3}A]$,*

$$\mathbb{E}\left(n^{1/3}|\hat{\lambda}_n(t) - \lambda(t)|\right)^p \leq K_{p,A}. \tag{7.12}$$

Proof. We denote $y_+ = \max(y, 0)$ and $y_- = -\min(y, 0)$ for all $y \in \mathbb{R}$. To go from $\hat{U}_n$ to $\hat{\lambda}_n$ we will make use of the following switch relation, that holds for all $t \in (0, 1]$ and $a \in \mathbb{R}$:

$$\hat{\lambda}_n(t) \geq a \Longleftrightarrow t \leq \hat{U}_n(a). \tag{7.13}$$

With $a_x = \lambda(t) + x$, it then follows from (7.7) and the switch relation (7.13) that

$$\begin{aligned}\mathbb{E}\left((\hat{\lambda}_n(t) - \lambda(t))_+\right)^p &= \int_0^\infty \mathbb{P}\left(\hat{\lambda}_n(t) - \lambda(t) \geq x\right) px^{p-1}dx \\ &= \int_0^\infty \mathbb{P}\left(\hat{U}_n(a_x) \geq t\right) px^{p-1}dx \\ &= I_1 + I_2 \qquad (7.14)\end{aligned}$$

where I_1 denotes the integral over $(0, \lambda(0) - \lambda(t)]$ while I_2 denotes the integral over $(\lambda(0) - \lambda(t), \infty)$. Consider I_1. Since $\lambda = \mu \circ F^{-1}$, it follows from the Taylor expansion that with $c = A_3/A_2$, we have $t - \lambda^{-1}(a_x) > cx$ for all $x \in (0, \lambda(0) - \lambda(t))$. Therefore, (7.9) implies that

$$\mathbb{P}\left(\hat{U}_n(a_x) \geq t\right) \leq \mathbb{P}\left(\hat{U}_n(a_x) - \lambda^{-1}(a_x) > cx\right) \leq K_1 \exp(-K_2 c^3 n x^3)$$

for all $x \in (0, \lambda(0) - \lambda(t))$. Hence,

$$I_1 \;\leq\; K_1 \int_0^{\lambda(0)-\lambda(t)} \exp(-K_2 c^3 n x^3) px^{p-1}dx \;\leq\; K_1 n^{-p/3} \int_0^\infty \exp(-K_2 c^3 y^3) py^{p-1}dy,$$

using the change of variable $y = n^{1/3}x$. The integral on the right hand side depends only on c and p, and is finite for all $p > 0$. Hence, with C_p/K_1 greater than this integral we obtain

$$I_1 \leq C_p n^{-p/3}. \qquad (7.15)$$

Now consider I_2. We have $a_x > \lambda(0)$ for all $x > \lambda(0) - \lambda(t)$ so it follows from (7.10) together with (7.9) (where $g(a_x) = 0$) that

$$\begin{aligned}I_2 \;&\leq\; K_1 \int_{\lambda(0)-\lambda(t)}^{2(\lambda(0)-\lambda(t))} \exp(-K_2 n t^3) px^{p-1}dx + K_1 \int_{2(\lambda(0)-\lambda(t))}^\infty \exp(-K_2(a_x - \lambda(0))^2 nt) px^{p-1}dx \\ &\leq\; K_1 \exp(-K_2 n t^3) 2^p (\lambda(0) - \lambda(t))^p + K_1 \int_{2(\lambda(0)-\lambda(t))}^\infty \exp(-K_2 x^2 nt/4) px^{p-1}dx,\end{aligned}$$

since $a_x - \lambda(0) \geq x/2$ for all $x \geq 2(\lambda(0) - \lambda(t))$. Since $\lambda = \mu \circ F^{-1}$, we have $|\lambda(t) - \lambda(0)| \leq A_2 t/A_3$ for all $t \in (0, 1]$ and therefore,

$$I_2 \;\leq\; K_1 2^p (A_2/A_3)^p \exp(-K_2 n t^3) t^p + K_1 (nt)^{-p/2} \int_0^\infty \exp(-K_2 y^2/4) py^{p-1}dy$$

using the change of variable $y = x\sqrt{nt}$. The function $t \mapsto \exp(-K_2 n t^3) t^p$ achieves its maximum on $[0, \infty)$ at the point $(3K_2 n/p)^{-1/3}$. This means that for all $t \geq 0$ we have

$$\exp(-K_2 n t^3) t^p \leq \exp(-p/3) \left(3K_2 n/p\right)^{-p/3}.$$

On the other hand, we have $(nt)^{-p/2} \leq A^{-p/2} n^{-p/3}$ for all $t \geq n^{-1/3}A$, where $A > 0$ is fixed. Combining this with the two preceding displays, we arrive at

$$I_2 \;\leq\; K_1 2^p (A_2/A_3)^p \exp(-p/3) \left(\frac{3K_2 n}{p}\right)^{-p/3} + K_1 A^{-p/2} n^{-p/3} \int_0^\infty \exp(-K_2 y^2/4) py^{p-1}dy$$

for all $t \geq n^{-1/3}A$, where the integral on the right hand side is finite. This means that there exists $K_{p,A} > 0$ such that $I_2 \leq K_{p,A} n^{-p/3}/2$ for all $t \geq n^{-1/3}A$. Combining this with (7.14) and (7.15) and possibly enlarging $K_{p,A} > 0$, we obtain

$$\mathbb{E}\left((\hat{\lambda}_n(t) - \lambda(t))_+\right)^p \leq K_{p,A} n^{-p/3}$$

for all $t \geq n^{-1/3}A$. It can be proved with similar arguments that the above inequality remains valid with $(\cdot)_+$ replaced by $(\cdot)_-$, and Lemma 7.4 follows. □

It is known that Grenander type estimators are inconsistent at the boundaries. However, the following lemma shows that such estimators remain bounded in the L_p-sense. The lemma, which is proved in the supplementary material, will be useful to go from Lemma 7.4 to Theorem 4.1.

LEMMA 7.5. *Assume (R4) and μ is nonincreasing with $|\mu(t)| \leq A_5$ for some $A_5 > 0$ and all $t \in [0, 1]$. Then, for all $p > 0$, there exists $K_1 > 0$ that depends only on p, A_5 and α, where α is taken from (R4), such that $\mathbb{E}|\hat{\lambda}_n(0)|^p \leq K_1$ and $\mathbb{E}|\hat{\lambda}_n(1)|^p \leq K_1$, $\forall n$.*

We are now in a position to prove the second assertion in Theorem 4.1. Since $\hat{\mu}_n$ is constant on all intervals $(X_{(i)}, X_{(i+1)}]$ for $i \in \{1, \dots, n-1\}$ and also on $[0, X_{(1)}]$, and F_n is constant on all intervals $[X_{(i)}, X_{(i+1)})$ for $i \in \{1, \dots, n-1\}$ and also on $[0, X_{(1)})$, it follows from (7.2) that for all $t \notin \{X_{(1)}, \dots, X_{(n)}\}$ we have $\hat{\mu}_n(t) = \hat{\lambda}_n(F_n(t) + n^{-1})$. But X has a continuous distribution so for a fixed t, we indeed have $t \notin \{X_{(1)}, \dots, X_{(n)}\}$ *a.s.*. Hence, for all $p \geq 1$ we have

$$\mathbb{E}\left((\hat{\mu}_n(t) - \mu(t))_+\right)^p \quad = \quad \mathbb{E}\left(\left(\hat{\lambda}_n(F_n(t) + n^{-1}) - \lambda(F(t))\right)_+\right)^p.$$

Using monotonicity of $\hat{\lambda}_n$, this means that

$$\begin{aligned}\mathbb{E}\left((\hat{\mu}_n(t) - \mu(t))_+\right)^p \quad \leq \quad & \mathbb{E}\left(\left(\hat{\lambda}_n(F(t) - n^{-1/2}\log n) - \lambda(F(t))\right)_+\right)^p \\ & +\mathbb{E}\left(\left(\hat{\lambda}_n(0) - \lambda(1)\right)_+^p \mathbb{1}_{F_n(t)+n^{-1} \leq F(t) - n^{-1/2}\log n}\right). \end{aligned} \tag{7.16}$$

It follows from the Hölder inequality that

$$\begin{aligned}&\mathbb{E}\left(\left(\hat{\lambda}_n(0) - \lambda(1)\right)_+^p \mathbb{1}_{F_n(t)+n^{-1} \leq F(t) - n^{-1/2}\log n}\right) \\ &\quad \leq \mathbb{E}^{1/2}\left(\left(\hat{\lambda}_n(0) - \lambda(1)\right)^{2p}\right) \mathbb{P}^{1/2}\left(F_n(t) + n^{-1} \leq F(t) - n^{-1/2}\log n\right) \\ &\quad \leq \mathbb{E}^{1/2}\left(\left(\hat{\lambda}_n(0) - \lambda(1)\right)^{2p}\right) \mathbb{P}^{1/2}\left(\sup_{t\in[0,1]} |F_n(t) - F(t)| > n^{-1/2}\log n\right).\end{aligned}$$

Combining this with Lemma 7.5 together with Corollary 1 in [22] yields

$$\mathbb{E}\left(\left(\hat{\lambda}_n(0) - \lambda(1)\right)_+^p \mathbb{1}_{F_n(t)+n^{-1} \leq F(t) - n^{-1/2}\log n}\right) \leq O(1)\left(2\exp(-2(\log n)^2)\right)^{1/2}$$

uniformly for all μ's satisfying the assumptions of the lemma. Hence, there exists C_p such that

$$\mathbb{E}\left(\left(\hat{\lambda}_n(0) - \lambda(1)\right)_+^p \mathbb{1}_{F_n(t)+n^{-1} \leq F(t) - n^{-1/2}\log n}\right) \leq C_p n^{-p/3} \tag{7.17}$$

for all $t \in [0,1]$. Now, consider the first term on the right hand side of (7.16). It follows from the convexity of the function $x \mapsto x^p$ that $(x+y)^p \le 2^{p-1}(x^p+y^p)$ for all positive numbers x and y. Therefore, with $t \ge n^{-1/3}$ and $x_n = F(t) - n^{-1/2}\log n$ we have

$$\begin{aligned}\mathbb{E}\left(\left(\hat\lambda_n(x_n) - \lambda(F(t))\right)_+\right)^p &\le 2^{p-1}\mathbb{E}\left(|\hat\lambda_n(x_n) - \lambda(x_n))|^p\right) + 2^{p-1}|\lambda(x_n) - \lambda(F(t)))|^p \\ &\le 2^{p-1}\mathbb{E}\left(|\hat\lambda_n(x_n) - \lambda(x_n))|^p\right) + 2^{p-1}(A_2/A_3)^p n^{-p/2}(\log n)^p\end{aligned}$$

since thanks to (4.3),

$$|\lambda(t) - \lambda(x)| \le A_2|t-x]/A_3 \text{ for all } t \neq x \in [0,1].$$

Let $A \le A_3/2$. For large n, we have $x_n \in [n^{-1/3}A, 1-n^{-1/3}A]$ for all $t \in [n^{-1/3}, 1-n^{-1/3}]$. Hence, the previous display combined with Lemma 7.4 ensures that there exists C_p such that

$$\mathbb{E}\left(\left(\hat\lambda_n(x_n) - \lambda(F(t))\right)_+\right)^p \le C_p n^{-p/3}$$

for all $t \in [n^{-1/3}, 1-n^{-1/3}]$ and n sufficiently large. Together with (7.17) and (7.16), this yields

$$\mathbb{E}\left((\hat\mu_n(t) - \mu(t))_+\right)^p \le 2C_p n^{-p/3}$$

for all $t \in [n^{-1/3}, 1-n^{-1/3}]$ and n sufficiently large. Possibly enlarging C_p, the previous inequality remains true for all n. To see this, suppose that the above display holds for all $n \ge n_{min}$. Now,

$$\mathbb{E}((\hat\mu_n(t) - \mu(t))_+)^p \le 2^{p-1}E(|\hat\mu_n(0)|^p \vee |\hat\mu_n(1)|^p) + 2^{p-1}|\mu(0)|^p \vee |\mu(1)|^p\,.$$

by monotonicity of both μ and $\hat\mu_n$, and using convexity of $x \mapsto x^p$. Hence, for $n < n_{min}$,

$$n^{p/3}\mathbb{E}((\hat\mu_n(t) - \mu(t))_+)^p \le (2^pK_1 + 2^pA_5)n_{min}^{p/3}\,,$$

where K_1 and A_5 are taken from Lemma 7.5. The negative part $\mathbb{E}\left((\hat\mu_n(t) - \mu(t))_-\right)^p$can be handled similarly, which completes the proof of Theorem 4.1. □

7.4.3. *Proof of Theorem 4.2.* Theorem 4.2 follows from Lemma 7.6 combined to Theorem 7.7 below since $\mu(1) = \lambda(1)$ and $\mu(0) = \lambda(0)$. Theorem 7.7 provides a precise bound for the bias of $\hat U_n$ whereas Lemma 7.6 makes the connection between the biases of $\hat\mu_n^{-1}$ and $\hat U_n$. The lemma is proved in the supplementary material, using that $\mu^{-1} = F^{-1}\circ g$ and $\hat\mu_n^{-1} = F_n^{-1}\circ \hat U_n$.

LEMMA 7.6. *Assume (R1), (R5), (R4). Let μ^{-1}, g be the generalized inverses of μ, λ. Then,*

$$\mathbb{E}\left(\hat\mu_n^{-1}(a) - \mu^{-1}(a)\right) = \frac{1}{f\circ F^{-1}(g(a))}\mathbb{E}\left(\hat U_n(a) - g(a)\right) + o(n^{-1/2})$$

where the small-o term is uniform in $a \in \mathbb{R}$.

THEOREM 7.7. *Assume (R1), (R5), (R3), (R4), v^2 has a bounded second derivative on $[0,1]$ and (4.4) holds for some $C>0$ and $s>1/2$. For an arbitrary constant $K>0$ we then have*

$$\mathbb{E}(\hat U_n(a)) - g(a) = o(n^{-1/2}) + O(n^{-(2s+3)/9}(\log n)^{25/2})$$

uniformly in $a \in \mathcal{J}_n := [\lambda(1) + Kn^{-1/6}\log n, \lambda(0) - Kn^{-1/6}\log n]$.

Proof. We first localize. For a given a we define

$$\hat{\hat{U}}_n(a) = \operatorname*{argmax}_{|u-g(b)|\le T_n n^{-1/3},\ u\in[0,1]} \{\Lambda_n(u) - au\} \tag{7.18}$$

with $T_n = n^\epsilon$ and b a random variable such that $b = a + O_p(n^{-1/2})$. Here, $\epsilon > 0$ is arbitrarily small. The variable b will be chosen in a convenient way later. Note that $\hat{\hat{U}}_n(a)$ is defined in a similar way as $\hat{U}_n(a)$, see (7.5), but with the location of the maximum taken on a shrinking neighborhood of $g(b)$ instead of being taken over the whole interval $[0,1]$. Although it may seem more natural to consider $b = a$, we will see that this choice is not the better one to derive precise bounds on the bias of $\hat{U}_n(a)$. For notational convenience, we do not make it explicit in the notation that $\hat{\hat{U}}_n(a)$ depends on b. The following lemma makes the connection between the bias of $\hat{U}_n(a)$ and that of the localized version; it is proved in the supplementary file.

LEMMA 7.8. *Assume (R1), (R2) and (R4). Let $a \in \mathbb{R}$ and b a random variable such that*

$$\mathbb{P}(|a-b| > x) \le K_1 \exp(-K_2 n x^2) \tag{7.19}$$

for all $x > 0$ where K_1, K_2 depend only on f, μ, σ. Then, $\mathbb{E}|\hat{U}_n(a) - \hat{\hat{U}}_n(a)| = o(n^{-1/2})$ uniformly.

In the sequel, we use the notation

$$L(t) = \int_0^t v^2 \circ F^{-1}(u)\, du \text{ for } t \in [0,1] \tag{7.20}$$

and the same notation $\mathcal{J}_n$ as in Theorem 7.7. We use L to normalize $\hat{\hat{U}}_n(a)$ in the following lemma, which is proved in the supplementary file. Thanks to the normalization with L, $\hat{\hat{U}}_n(a)$ can be approached by the location of the maximum of a drifted Brownian motion, see (7.27).

LEMMA 7.9. *Assume (R1), (R5), (R3), (R4). Let $a \in \mathcal{J}_n$ and b as in (7.19) for all $x > 0$, where K_1, K_2 depend only on f, μ and v. Assume, furthermore, that $\mathbb{E}(b) = a + o(n^{-1/2})$ and that v^2 and μ have a continuous first derivative on $[0,1]$. Uniformly in $a \in \mathcal{J}_n$, we then have*

$$\mathbb{E}(\hat{\hat{U}}_n(a) - g(a)) = \mathbb{E}\left(\frac{L(\hat{\hat{U}}_n(a)) - L(g(b))}{L'(g(a))}\right) + o(n^{-1/2})$$

With B_n and L taken from (8.5) and (7.20) respectively, let

$$\phi_n(t) = \frac{L''(t)}{\sqrt{n}L'(t)} B_n(t) \tag{7.21}$$

Moreover, let A_n be the event that all inequalities in (7.22) and (7.23) below hold true :

$$\sup_{u\in[0,1]} |B_n(u)| \le \log n, \qquad \sup_{|u-v|\le T_n n^{-1/3}\sqrt{\log n}} |B_n(u) - B_n(v)| \le \sqrt{T_n} n^{-1/6} \log n, \tag{7.22}$$

$$\sup_{u\in[0,1]} \left| F_n^{-1}(u) - F^{-1}(u) - \frac{1}{\sqrt{n} f(F^{-1}(u))} B_n(u) \right| \le n^{\delta-1}, \tag{7.23}$$

where $\delta \in (0, 1/3)$ can be chosen as small as we wish. We will prove below that $\mathbb{P}(A_n) \to 1$ as $n \to \infty$, see (7.33). The following lemma is proved in the supplement. Here and in the sequel,

$$\Lambda(t) = \int_0^t \lambda(u) du. \tag{7.24}$$

Lemma 7.10. *Let $q > 0$, $a \in \mathcal{J}_n$ and*

$$b = a - \frac{B_n(g(a))}{\sqrt{n}} \lambda'(g(a)). \tag{7.25}$$

Under the assumptions of Theorem 7.7, on A_n, conditionally on $(X_1, \ldots, X_n)$, the variable

$$n^{1/3}(L(\hat{\hat{U}}_n(a)) - L(g(b))) \tag{7.26}$$

has the same distribution as

$$\operatorname*{argmax}_{u \in I_n(b)} \{D_n(b, u) + W_{g(b)}(u) + R_n(a, b, u)\}, \tag{7.27}$$

where for all $t \in [0, 1]$,

$$W_t(u) = \frac{n^{1/6}}{\sqrt{1 + \phi_n(t)}} \left[W_n \left(L_n(t) + n^{-1/3} u (1 + \phi_n(t) \right) - W_n(L_n(t)) \right], \ u \in \mathbb{R}, \tag{7.28}$$

with W_n being a standard Brownian motion under $\mathbb{P}^X$,

$$I_n(b) = \left[n^{1/3} \left(L(g(b) - n^{-1/3} T_n) - L(g(b)) \right), n^{1/3} \left(L(g(b) + n^{-1/3} T_n) - L(g(b)) \right) \right],$$

$$D_n(b, u) = n^{2/3} \left(\Lambda \circ L^{-1}(L(g(b)) + n^{-1/3} u) - \Lambda(g(b)) - b L^{-1}(L(g(b)) + n^{-1/3} u) + b g(b) \right),$$

and with $T_n = n^\epsilon$ for some sufficiently small $\epsilon > 0$,

$$\mathbb{P}^X \left(\sup_{u \in I_n(b)} |R_n(a, b, u)| > x \right) \leq K_q x^{-q} n^{1 - q/3} \tag{7.29}$$

for all $x > 0$, where $K_q > 0$ does not depend on n.

It follows from Lemma 7.10 that conditionally on $(X_1, \ldots, X_n)$, on A_n the variable in (7.26) has the same expectation as the variable defined in (7.27). The following lemma, which is proved in the supplementary file, shows that R_n is negligible in (7.27) in the sense that this expectation, up to a negligible remainder term, is equal to the expectation of the variable

$$V_n(b) = \operatorname*{argmax}_{|u| \leq (L'(g(b)))^{4/3} \log n} \{D_n(b, u) + W_{g(b)}(u)\}.$$

Lemma 7.11. *Let $a \in \mathcal{J}_n$ and let b be given by (7.25). Under the assumptions of Theorem 7.7, with $T_n = n^\epsilon$ for some sufficiently small $\epsilon > 0$, there exists $K > 0$ such that on A_n, we have*

$$\left| \mathbb{E}^X \left(n^{1/3} (L(\hat{\hat{U}}_n(a)) - L(g(b))) \right) - \mathbb{E}^X(V_n(b)) \right| \leq K n^{-1/6} L'(g(b)) (\log n)^{-1}.$$

The following lemma is proved in the supplementary file.

LEMMA 7.12. *Assume (R1), (R5), (R3). Assume, furthermore, that v^2 has a bounded second derivative on $[0,1]$ and (4.4) holds for some $C>0$ and $s>1/2$. Let $a\in\mathcal{J}_n$ and b be given by (7.25). With $T_n=n^\epsilon$ for some small enough $\epsilon>0$, there exists $K>0$ such that on A_n, we have*

$$\left|\mathbb{E}^X(V_n(b))\right|\le Kn^{-2s/9}L'(g(b))(\log n)^{25/2}.$$

We are now in a position to prove Theorem 7.7. Let $a\in\mathcal{J}_n$ and let $\hat{\hat{U}}_n(a))$ be defined by (7.18) where b is taken from (7.25). Since λ' is bounded, there exists $K>0$ such that

$$\mathbb{P}(|a-b|>x)\le\mathbb{P}\left(\sup_{u\in[0,1]}|B_n(u)|>Kx\sqrt{n}\right)\quad\text{for all } x>0.$$

Then, with the representation $B_n(u)=W(u)-uW(1)$ in distribution of processes, where W is a standard Brownian motion, we conclude from the triangle inequality that

$$\mathbb{P}(|a-b|>x)\le\mathbb{P}\left(\sup_{u\in[0,1]}|W(u)|>Kx\sqrt{n}/2\right)=2\mathbb{P}\left(\sup_{u\in[0,1]}W(u)>Kx\sqrt{n}/2\right).$$

For the last equality, we used symmetry of W. Then, it follows from [24, Proposition 1.8] that (7.19) holds for all $x>0$, where $K_1=2$ and K_2 depends only on λ. By lemma 7.8, we then have

$$\mathbb{E}(\hat{U}_n(a)-g(a))=\mathbb{E}(\hat{\hat{U}}_n(a)-g(a))+o(n^{-1/2})$$

where the small-o term is uniform in $a\in\mathcal{J}_n$. Since B_n is a centered process, we have $\mathbb{E}(b)=a$, so Lemma 7.9 combined with the preceding display ensures that

$$\mathbb{E}(\hat{U}_n(a)-g(a))=\mathbb{E}\left(\frac{L(\hat{\hat{U}}_n(a))-L(g(b))}{L'(g(a))}\right)+o(n^{-1/2})\tag{7.30}$$

uniformly in $a\in\mathcal{J}_n$. Now, conditionally on $(X_1,\dots,X_n)$, on A_n we have

$$\left|\mathbb{E}^X\left(n^{1/3}(L(\hat{\hat{U}}_n(a))-L(g(b)))\right)-\mathbb{E}^X(V_n(b))\right|\le K_3n^{-1/6}L'(g(b))(\log n)^{-1}$$

and

$$\left|\mathbb{E}^X(V_n(b))\right|\le K_3n^{-2s/9}L'(g(b))(\log n)^{25/2}.$$

Here, we use Lemma 7.11 and Lemma 7.12 with A_n being the event that all inequalities in (7.22) and (7.23) hold true. It then follows from the triangle inequality that

$$\mathbb{E}\left(\left|\mathbb{E}^X\left(n^{1/3}(L(\hat{\hat{U}}_n(a))-L(g(b)))\right)\right|\mathbb{1}_{A_n}\right)\le K_3\mathbb{E}(L'(g(b)))\beta_n$$

where $\beta_n=n^{-2s/9}(\log n)^{25/2}+n^{-1/6}(\log n)^{-1}$. But $L'\circ g$ is a Lipschitz function, so we have

$$\mathbb{E}\left|L'(g(b))-L'(g(a))\right|\le K_4\mathbb{E}|b-a|\le K_5n^{-1/2},$$

using (8.28) together with the Jensen inequality for the last inequality. Using (8.24) and the two previous displays yields

$$\mathbb{E}\left(\left|\mathbb{E}^X\left(n^{1/3}(L(\hat{\hat{U}}_n(a))-L(g(b)))\right)\right|\mathbb{1}_{A_n}\right)\le 2K_3L'(g(a))\beta_n\tag{7.31}$$

for n sufficiently large. On the other hand, denoting by $\bar{A}_n$ the complementary of A_n, it follows from the Hölder inequality together with the Jensen inequality that

$$\begin{aligned}\mathbb{E}\left(\left|\mathbb{E}^X\left(n^{1/3}(L(\hat{\hat{U}}_n(a))-L(g(b)))\right)\right|\mathbb{1}_{\bar{A}_n}\right)\\ \leq \mathbb{E}^{1/2}\left(n^{1/3}(L(\hat{\hat{U}}_n(a))-L(g(b)))\right)^2\mathbb{P}^{1/2}(\bar{A}_n).\end{aligned}\tag{7.32}$$

Then, we derive from (8.26) and (8.28) that the expectation on the right-hand side is finite. Now, consider $\mathbb{P}(\bar{A}_n)$ on the right-hand side. It follows from the Markov inequality together with Lemma 8.2 that for all $r \geq 1$ we have

$$\begin{aligned}\mathbb{P}\left(\sup_{u\in[0,1]}\left|F_n^{-1}(u)-F^{-1}(u)-\frac{B_n(u)}{\sqrt{n}f(F^{-1}(u))}\right|>n^{\delta-1}\right) &\leq K_6\,(\log n)^r\,n^{-r\delta}\\ &\leq K_6\left(n^{-1/6}L'(g(a))(\log n)^{-1}\right)^2\end{aligned}$$

for large n, provided that $r > 2/(3\delta)$. The Brownian motion satisfies the assumption (A2) with $\tau = 1$ of Lemma 5.1 in [12] (see the proof of Corollary 3.1 in that paper), so we conclude that

$$\mathbb{P}^{1/2}(\bar{A}_n) \leq K_7 n^{-1/6}L'(g(a))(\log n)^{-1}\tag{7.33}$$

for n sufficiently large. Hence, (7.32) yields

$$\mathbb{E}\left(\left|\mathbb{E}^X\left(n^{1/3}(L(\hat{\hat{U}}_n(a))-L(g(b)))\right)\right|\mathbb{1}_{\bar{A}_n}\right) \leq K_8 n^{-1/6}L'(g(a))(\log n)^{-1}.$$

Together with (7.31), this yields

$$\mathbb{E}\left(\left|\mathbb{E}^X\left(n^{1/3}(L(\hat{\hat{U}}_n(a))-L(g(b)))\right)\right|\right) \leq K_9 L'(g(a))\beta_n.$$

Hence, with the Jensen inequality we arrive at

$$\mathbb{E}\left(\frac{n^{1/3}(L(\hat{\hat{U}}_n(a))-L(g(b)))}{L'(g(a))}\right) = O(\beta_n).$$

Combining this with (7.30) completes the proof of Theorem 7.7. □

7.4.4. *Proof of Theorem 4.3.* The following lemma is proved in the supplementary file.

LEMMA 7.13. *Assume (R1), (R2), (R4). With $K > 0$ arbitrary, there exists positive K_1, K_2 with*

$$\mathbb{P}\left(|\hat{\mu}_n(t)-\mu(t)| > n^{-1/3}\log n\right) \leq K_1\exp(-K_2(\log n)^3)\tag{7.34}$$

for all $t \in [Kn^{-1/6}\log n, 1-Kn^{-1/6}\log n]$, and

$$\mathbb{E}\left(\hat{\mu}_n(t)-\mu(t)\right) = \mathbb{E}\left[\left(\hat{\mu}_n(t)-\mu(t)\right)\mathbb{1}_{|\hat{\mu}_n(t)-\mu(t)|\leq n^{-1/3}\log n}\right] + o(n^{-1/2})$$

where the small-o term is uniform in $t \in [Kn^{-1/6}\log n, 1-Kn^{-1/6}\log n]$.

We turn to the proof of Theorem 4.3. Distinguishing the positive and negative parts of $\hat{\mu}_n(t)-\mu(t)$, we derive from (7.7) together with Lemma 7.13 that $\mathbb{E}\left(\hat{\mu}_n(t)-\mu(t)\right)=I_1-I_2+o(n^{-1/2})$ where

$$I_1=\int_0^{n^{-1/3}\log n}\mathbb{P}\left(\hat{\mu}_n(t)-\mu(t)\geq x\right)dx \text{ and } I_2=\int_0^{n^{-1/3}\log n}\mathbb{P}\left(\mu(t)-\hat{\mu}_n(t)>x\right)dx.$$

Consider I_1. Since $\hat{\mu}_n^{-1}=F_n^{-1}\circ\hat{U}_n$, it follows from the switch relation and (8.1) that

$$\begin{aligned}
I_1 &= \int_0^{n^{-1/3}\log n}\mathbb{P}\left(\hat{\mu}_n^{-1}(x+\mu(t))\geq t\right)dx\\
&= \int_0^{n^{-1/3}\log n}\mathbb{P}\left(F^{-1}\circ\hat{U}_n(x+\mu(t))\geq t-O\left(n^{-1/2}\log n\right)\right)dx+o(n^{-1/2})\\
&= \int_0^{n^{-1/3}\log n}\mathbb{P}\left(\hat{U}_n(x+\mu(t))\geq F(t)-O(n^{-1/2}\log n)\right)dx+o(n^{-1/2}),
\end{aligned}$$

where the small o-term is uniform in $t\in[c_1,c_2]$. We have $g\circ\mu=F$ and $g'\circ\mu=(\lambda'\circ F)^{-1}$ so it follows from the Taylor expansion that

$$g(x+\mu(t))=F(t)-\frac{x}{|\lambda'\circ F(t)|}+O(x^{1+s})$$

for all $t\in[c_1,c_2]$ and $x\in[0,n^{-1/3}\log n]$, where s is taken from (4.4) and c_1,c_2 are as in the statement of the theorem. Since $x^{1+s}\leq n^{-1/2}\log n$ for all $x\leq n^{-1/3}\log n$ for large n, we conclude that

$$I_1 = \int_0^{n^{-1/3}\log n}\mathbb{P}\left(\hat{U}_n(a_x)-g(a_x)>\frac{x}{|\lambda'\circ F(t)|}-O(n^{-1/2}\log n)\right)dx+o(n^{-1/2}),$$

uniformly, where we set $a_x=\mu(t)+x$. But it follows from (7.5) together with (7.18) that

$$\mathbb{P}\left(\hat{\hat{U}}_n(a_x)\neq\hat{U}_n(a_x)\right) \leq \mathbb{P}\left(|\hat{U}_n(a_x)-g(b_x)|>T_nn^{-1/3}\right) \tag{7.35}$$

for all $x>0$, where we recall that $T_n=n^{\epsilon}$ for some arbitrarily small $\epsilon>0$, and b_x satisfies (7.19) with a replaced by a_x. Together with Lemma 7.2, this yields

$$I_1 = \int_0^{n^{-1/3}\log n}\mathbb{P}\left(\hat{\hat{U}}_n(a_x)-g(a_x)>\frac{x}{|\lambda'\circ F(t)|}-O(n^{-1/2}\log n)\right)dx+o(n^{-1/2}),$$

uniformly in t. Using again (7.35) and Lemma 7.2, we then derive from (8.25) in the supplementary file that

$$I_1 = \int_0^{n^{-1/3}\log n}\mathbb{P}\left(\frac{L(\hat{\hat{U}}_n(a_x))-L(g(b_x))}{L'(g(a_x))}>\frac{x}{|\lambda'\circ F(t)|}-O(n^{-1/2}\log n)\right)dx+o(n^{-1/2}),$$

where b_x is given by (7.25) with a replaced by a_x and B_n being taken from Lemma 8.2. Since $L'\circ g=v^2\circ\mu^{-1}$, we have

$$\mathbb{P}(L'(g(b_x))\leq c_0\gamma)\leq\mathbb{P}(\mu^{-1}(b_x)\leq\gamma)+\mathbb{P}(1-\mu^{-1}(b_x)\leq\gamma)$$

for all $\gamma > 0$ and $x \in (0, n^{-1/3}\log n)$, where c_0 is taken from (R3). Consider the first probability on the right-hand side. Assume that $\gamma > 0$ is chosen small enough so that $c_1 > \gamma$. By monotonicity of μ and the definition of b_x, there exists a positive constant K_1 such that for $x \in (0, n^{-1/3}\log n]$ we have

$$\begin{aligned}\mathbb{P}(\mu^{-1}(b_x) \leq \gamma) &\leq \mathbb{P}\left(\mu(t) + x - \frac{B_n(g(x_a))}{\sqrt{n}}\lambda'(g(a_x) \geq \mu(\gamma)\right)\\ &\leq \mathbb{P}\left(|B_n(g(x_a))| \geq K_1\sqrt{n}(c_1-\gamma)\right) \leq 4\exp(-K_1^2 n(c_1-\gamma)^2/2).\end{aligned}$$

It can be proved likewise that $\mathbb{P}(1-\mu^{-1}(b_x) \leq \gamma) \leq 4\exp(-K_1^2 n(1-c_2-\gamma)^2/2)$ provided $\gamma > 0$ is chosen sufficiently small so that $c_2 + \gamma < 1$. Hence, we can restrict attention to the event $\{L'(g(b_x)) > c_0\gamma\}$, which mean that $L'(g(b_x)$ cannot go to zero. Then, using (8.39) with $\delta = n^{1/3}\gamma_n$ for some $\gamma_n \in (n^{-1/2}\log n, n^{-1/3}\log n)$ to be chosen later, we have

$$\begin{aligned}I_1 &= \int_0^{n^{-1/3}\log n} \mathbb{E}\mathbb{P}^X\left(\frac{n^{-1/3}V_n(b_x)}{L'(g(a_x))} > \frac{x}{|\lambda'\circ F(t)|} - O(\gamma_n)\right)dx + o(n^{-1/2})\\ &\quad + O\left(n^{-1/3}(\log n)^2 n^{(3-q)/(3(q+1))}(n^{1/3}\gamma_n)^{-3q/(2(q+1))}\right)\end{aligned}$$

where q can be chosen arbitrarily large. For arbitrary $\phi > 0$ we can choose q large enough so that

$$\begin{aligned}I_1 &= \int_0^{n^{-1/3}\log n} \mathbb{E}\mathbb{P}^X\left(\frac{n^{-1/3}V_n(b_x)}{L'(g(a_x))} > \frac{x}{|\lambda'\circ F(t)|} - O(\gamma_n)\right)dx + o(n^{-1/2})\\ &\quad + O(n^{-7/6+\phi}\gamma_n^{-3/2-\phi})\\ &= \int_0^{n^{-1/3}\log n} \mathbb{E}\mathbb{P}^X\left(n^{-1/3}V_n(b_x) > \frac{xv^2(t)}{|\lambda'\circ F(t)|} - O(\gamma_n)\right)dx + o(n^{-1/2})\\ &\quad + O(n^{-7/6+\phi}\gamma_n^{-3/2-\phi}).\end{aligned}$$

Now, using (8.42) in the supplementary file with $s = 1$ and $\delta = n^{1/3}\gamma_n$ proves that I_1 is equal to

$$\int_0^{n^{-1/3}\log n} \mathbb{E}\mathbb{P}^X\left(n^{-1/3}V(b_x) > \frac{xv^2(t)}{|\lambda'\circ F(t)|} - O(\gamma_n)\right)dx + o(n^{-1/2}) + O(n^{-7/6+\phi}\gamma_n^{-3/2-\phi}).$$

Recall that $g\circ\mu = F$. Let $Z(t) = \operatorname{argmax}_{u\in\mathbb{R}}\{-d(F(t))u^2 + W(u)\}$, where $d = |\lambda'|/(2(L')^2)$ and W is a standard Brownian motion. Under $\mathbb{P}^X$, $Z(t)$ has the same law as the location of the maximum of $u \mapsto -d(F(t))u^2 + W_{g(b_x)}(u)$ on $\mathbb{R}$. On the event $\{\sup_{t\in[0,1]}|B_n(t)| \leq \log n\}$,

$$V(b_x) = \underset{|u|\leq(L'(g(b_x)))^{4/3}\log n}{\operatorname{argmax}}\{-d(F(t))u^2 + W_{g(b_x)}(u) + R_n(u,x,t)\}$$

where

$$\sup_{|u|\leq(L'(g(b_x)))^{4/3}\log n}|R_n(u,x,t)| = O(n^{-s/3}(\log n)^{2+s})$$

uniformly in $t \in [c_1, c_2]$ and $x \in (0, n^{-1/3}\log n)$. It then follows from Proposition 1 in [10] (see also the comments just above this proposition) that there are versions of $Z(t)$ and $V(b_x)$, and constants $K_1, K_2, K_3 > 0$, such that on $\{\sup_{t\in[0,1]}|B_n(t)| \leq \log n\}$ and for large n, we have

$$\begin{aligned}\mathbb{P}^X\left(|V(b_x) - Z(t)| > n^{1/3}\gamma_n\right) &\leq \mathbb{P}^X\left(2\sup_{|u|\leq(L'(g(b_x)))^{4/3}\log n}|R_n(u,x,t)| > x(n^{1/3}\gamma_n)^{3/2}\right)\\ &\quad + K_1 x\log n + 2\mathbb{P}^X(|Z(t)| > K_2\log n)\end{aligned}$$

where $x = K_3(n^{1/3}\gamma_n)^{-3/2}n^{-s/3}(\log n)^{2+s}$. With large K_3, the probability on the right hand side is equal to zero. Hence, there exists $K_4 > 0$ such that on $\{\sup_{t\in[0,1]}|B_n(t)| \le \log n\}$ we have

$$\begin{aligned}\mathbb{P}^X\left(|V(b_x) - Z(t)| > n^{1/3}\gamma_n\right) &\le K_4(n^{1/3}\gamma_n)^{-3/2}n^{-s/3}(\log n)^{3+s} + 2\mathbb{P}^X\left(|Z(t)| > K_2\log n\right)\\ &\le K_4(n^{1/3}\gamma_n)^{-3/2}n^{-s/3}(\log n)^{3+s} + 4\exp(-K_5(\log n)^3)\end{aligned}$$

for some $K_5 > 0$. For the last inequality, we used [10, Theorem 4]. The second term on the right hand side is negligible as compared to the first one, so there exists $K_6 > 0$ such that

$$\mathbb{P}^X\left(|V(b_x) - Z(t)| > n^{1/3}\gamma_n\right) \le K_6(n^{1/3}\gamma_n)^{-3/2}n^{-s/3}(\log n)^{3+s}.$$

Since $s = 1$, we obtain that I_1 is equal to

$$(7.36)\int_0^{n^{-1/3}\log n}\mathbb{P}\left(n^{-1/3}Z(t) > \frac{xv^2(t)}{|\lambda'\circ F(t)|} - O(\gamma_n)\right)dx + o(n^{-1/2}) + O(n^{-7/6+\phi}\gamma_n^{-3/2-\phi}).$$

Consider the integral on the right-hand side. There exists $K > 0$ such that the integral on the right hand side of (7.36) is bounded from above by

$$\begin{aligned}&\int_0^{n^{-1/3}\log n}\mathbb{P}\left(n^{-1/3}Z(t) > \frac{(x-K\gamma_n)v^2(t)}{|\lambda'\circ F(t)|}\right)dx\\ &\le \int_0^{n^{-1/3}\log n}\mathbb{P}\left(n^{-1/3}Z(t) > \frac{yv^2(t)}{|\lambda'\circ F(t)|}\right)dy + O(\gamma_n)\end{aligned}$$

using the change of variable $y = x - K\gamma_n$. Similarly, the integral in (7.36) is bounded below by

$$\int_0^{n^{-1/3}\log n}\mathbb{P}\left(n^{-1/3}Z(t) > \frac{yv^2(t)}{|\lambda'\circ F(t)|}\right)dy + O(\gamma_n)$$

and therefore,

$$I_1 = \int_0^{n^{-1/3}\log n}\mathbb{P}\left(n^{-1/3}Z(t) > \frac{xv^2(t)}{|\lambda'\circ F(t)|}\right)dx + O(\gamma_n) + O(n^{-7/6+\phi}\gamma_n^{-3/2-\phi}).$$

Choose γ_n that approximately realize the best trade-of between the two big-O-terms, that is such that $\gamma_n = n^{-7/6}\gamma_n^{-3/2}$. Then $\gamma_n = n^{-7/15}$, we conclude that for arbitrarily small $\phi > 0$,

$$I_1 = \int_0^{n^{-1/3}\log n}\mathbb{P}\left(n^{-1/3}Z(t) > \frac{xv^2(t)}{|\lambda'\circ F(t)|}\right)dx + O(n^{-14/30+\phi}).$$

With similar arguments, we obtain that for arbitrarily small $\phi > 0$,

$$I_2 = \int_0^{n^{-1/3}\log n}\mathbb{P}\left(n^{-1/3}Z(t) < -\frac{xv^2(t)}{|\lambda'\circ F(t)|}\right)dx + O(n^{-7/15+\phi}).$$

But $Z(t)$ has the same distribution as $-Z(t)$ for all t so the two preceding displays yield that $I_1 - I_2 = O(n^{-7/15+\phi})$. This completes the proof of Theorem 4.3. □

Acknowledgements. The authors thank Arijit Chakrabarty, Guang Cheng, Subhashis Ghoshal, Piet Groeneboom and Jon Wellner for some illuminating discussions and correspondences. They also express their gratitude to the Associate Editor and two referees whose comments and suggestions have led to a substantially improved version of the paper.

References.

[1] Banerjee, M. (2005). Likelihood ratio tests under local and fixed alternatives in monotone function problems. *Scand. J. Statist.*, 32(4):507–525.

[2] Banerjee, M. (2007). Likelihood based inference for monotone response models. *Ann. Statist.*, 35(3):931–956.

[3] Banerjee, M. (2008). Estimating monotone, unimodal and u–shaped failure rates using asymptotic pivots. *Statistica Sinica*, pages 467–492.

[4] Banerjee, M. and McKeague, I. W. (2007). Confidence sets for split points in decision trees. *Ann. Statist.*, 35(2):543–574.

[5] Banerjee, M. and Wellner, J. A. (2005). Confidence intervals for current status data. *Scand. J. Statist.*, 32(3):405–424.

[6] Brown, L. D., Low, M. G., and Zhao, L. H. (1997). Superefficiency in nonparametric function estimation. *Ann. Statist.*, 25(6):2607–2625.

[7] Brunk, H. D. (1955). Maximum likelihood estimates of monotone parameters. *Ann. Math. Statist.*, 26:607–616.

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

[9] Chernoff, H. (1964). Estimation of the mode. *Ann. Inst. Statist. Math.*, 16:31–41.

[10] Durot, C. (2002). Sharp asymptotics for isotonic regression. *Probability theory and related fields*, 122(2):222–240.

[11] Durot, C. (2008). Monotone nonparametric regression with random design. *Math. Methods Statist.*, 17(4):327–341.

[12] Durot, C. and Lopuhaä, H. P. (2014). A kiefer-wolfowitz type of result in a general setting, with an application to smooth monotone estimation. *Electronic Journal of Statistics*, 8(2):2479–2513.

[13] Durot, C. and Thiébot, K. (2006). Bootstrapping the shorth for regression. *ESAIM: Probability and Statistics*, 10:216–235.

[14] Grenander, U. (1956). On the theory of mortality measurement, part ii. *Skand. Akt.*, 39:125–153.

[15] Groeneboom, P. and Jongbloed, G. (2014). *Nonparametric Estimation under Shape Constraints: Estimators, Algorithms and Asymptotics*, volume 38. Cambridge University Press.

[16] Groeneboom, P. and Wellner, J. A. (1992a). *Information bounds and nonparametric maximum likelihood estimation*, volume 19 of *DMV Seminar*. Birkhäuser Verlag, Basel.

[17] Groeneboom, P. and Wellner, J. A. (1992b). *Information bounds and nonparametric maximum likelihood estimation*, volume 19. Springer Science & Business Media.

[18] Huang, J. and Wellner, J. A. (1995). Estimation of a monotone density or monotone hazard under random censoring. *Scand. J. Statist.*, pages 3–33.

[19] Kim, J. and Pollard, D. (1990). Cube root asymptotics. *Annals of Statistics*, 18:191–219.

[20] Li, R., Lin, D. K., and Li, B. (2013). Statistical inference in massive data sets. *Applied Stochastic Models in Business and Industry*, 29(5):399–409.

[21] Manski, C. F. (1975). Maximum score estimation of the stochastic utility model of choice. *J. Econometrics*, 3(3):205–228.

[22] Massart, P. (1990). The tight constant in the dvoretzky-kiefer-wolfowitz inequality. *The Annals of Probability*, pages 1269–1283.

[23] Prakasa Rao, B. (1969). Estimation of a unimodal density. *Sankhyā Ser. A*, 31:23–36.

[24] Revuz, D. and Yor, M. (2013). *Continuous martingales and Brownian motion*, volume 293. Springer Science & Business Media.

[25] 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. John Wiley & Sons Ltd., Chichester.

[26] Rousseeuw, P. J. (1984). Least median of squares regression. *J. Amer. Statist. Assoc.*, 79(388):871–880.

[27] Tsybakov, A. B. (2009). *Introduction to nonparametric estimation.* Springer Series in Statistics. Springer, New York. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats.

[28] Wright, F. T. (1981). The asymptotic behavior of monotone regression estimates. *Ann. Statist.*, 9(2):443–448.

[29] Zhang, Y., Duchi, J., and Wainwright, M. (2013). Divide and conquer kernel ridge regression. In *Conference on Learning Theory*, pages 592–617.

[30] Zhao, T., Cheng, G., and Liu, H. (2014). A partially linear framework for massive heterogeneous data. *arXiv preprint arXiv:1410.8570.*

University of Michigan
451, West Hall,
1085 South University
Ann Arbor, MI 48109
E-mail: moulib@umich.edu

Université Paris Nanterre
200 avenue de la république
92000 Nanterre, France
E-mail: cecile.durot@gmail.com

Columbia University
1255 Amsterdam Av.
Room # 1032 SSW
New York, NY 10027
E-mail: bodhi@stat.columbia.edu