Electronic Journal of Statistics
Vol. 13 (2019) 1926–1977
ISSN: 1935-7524
https://doi.org/10.1214/19-EJS1559

Circumventing superefficiency: An effective strategy for distributed computing in non-standard problems

Moulinath Banerjee* **and Cécile Durot†**

University of Michigan, 275, West Hall, 1085 South University Ann Arbor, MI 48109
and
Modal'x, Université Paris Nanterre, F-92001, Nanterre, France
e-mail: moulib@umich.edu; cecile.durot@parisnanterre.fr

Abstract: We propose a strategy for computing estimators in some non-standard M-estimation problems, where the data are distributed across different servers and the observations across servers, though independent, can come from heterogeneous sub-populations, thereby violating the identically distributed assumption. Our strategy fixes the super-efficiency phenomenon observed in prior work on distributed computing in (i) the isotonic regression framework, where averaging several isotonic estimates (each computed at a local server) on a central server produces super-efficient estimates that do not replicate the properties of the global isotonic estimator, i.e. the isotonic estimate that would be constructed by transferring *all the data* to a single server, and (ii) certain types of M-estimation problems involving optimization of discontinuous criterion functions where M-estimates converge at the cube-root rate. The new estimators proposed in this paper work by smoothing the data on each local server, communicating the smoothed summaries to the central server, and then solving a non-linear optimization problem at the central server. They are shown to replicate the asymptotic properties of the corresponding global estimators, and also overcome the super-efficiency phenomenon exhibited by existing estimators.

AMS 2000 subject classifications: Primary 62G05, 62G20, 62G08; secondary62E20.
Keywords and phrases: Cube-root asymptotics, distributed computing, isotonic regression, local minimax risk, superefficiency.

Received June 2018.

Contents

*NSF grant DMS-1712962.
†Labex MME-DII (ANR11-LBX-0023-01).

1. Background

Distributed computing has now become significant in the practice of statistics as well as other branches of data science. Large volumes of data, often relating to the same or closely related studies or experiments, are no longer stored on one single computer; rather, they are distributed across a number of platforms in some structured manner, owing partly to natural memory constraints on individual machines, and partly for convenience. This, typically, poses problems for computing optimal estimates of parameters of interest from the data at hand. Conventional statistical estimates are generally obtained under the premise that the totality of the data is accessible to a single computing device and can be processed at one stroke, yielding estimates that are *optimal* in some quantitatively defined sense. However, this is not automatically the case in a distributed environment. The calculation of *global estimates* that require simultaneous processing of all available data then entails transferring the entire bulk of data from different computers to a central machine, which in itself can be both time and resource consuming, followed by a potentially complex computation on the aggregated data (of massive volume), which may be infeasible under many circumstances.

Divide and conquer algorithms are a standard approach to addressing these issues in a distributed computing environment. The idea behind this is as follows: suppose the entire data set is stored across a number of machines. On each machine, calculate a natural estimate of the parameter of interest from the data on it and transfer this estimate to a central machine. Next, combine the estimators thus obtained, at the central machine in a judicious way to produce a final estimate, the so-called *pooled estimate*, which replicates the properties of the natural *global estimate*, i.e. the one we could have computed were it feasible

to store and analyze all available data on one machine. The term 'replicates the properties' can be understood in various ways and is often specific to the problem at hand: one might be able to show that the pooled and the global estimates have the same rate of convergence, or are comparable, up to constants, in terms of a certain measure of risk, or it might even be possible to demonstrate that the pooled estimate and the global estimate have the same limit distributions under appropriate conditions. The other important factor is computational burden: one would expect that the divide and conquer algorithm is not substantially more computationally onerous than the global estimator. As the literature on distributed computing is enormous, here we provide a selection of instances of research on distributed computing problems in a variety of statistical/machine-learning contexts: see, e.g. [10], [12], [26], [27], [6], [19], [24]. The above papers illustrate that the sample splitting approach buys computational dividends, yet statistical optimality [in the sense that the resulting estimator is as efficient (or minimax rate optimal) as the global estimate based on applying the estimation algorithm to the entire data set] is retained.

In the nonparametric function estimation context, most results of the divide and conquer (DC) type focus on estimation under smoothness constraints, where the essence of the strategy is to compute a smoothed estimator of the unknown function at each server and combine the estimators at the central server, by averaging; this strategy is employed, for example, in [12], [26], [27]. However, the averaging strategy leads to highly problematic pooled estimators in non-regular function estimation problems, e.g. function estimation under a monotonicity constraint, where the least squares estimates under the monotonicity constraint are non-standard/non-regular in the sense that they are non-linear in the data, and have non-Gaussian limit distributions. This is the core content of the recent work by [3] [henceforth BDS] where it is demonstrated that in monotone function estimation, the 'pooled-by-averaging' estimator [henceforth, generally referred to as BDSE] becomes *super-efficient*: its ARE (asymptotic relative efficiency in terms of MSE) with respect to the global monotone least squares estimator computed at any single model goes to infinity, whereas, in the uniform sense, the ARE goes to 0, i.e. the maximal MSE of BDSE over a collection of models relative to that of the global least squares estimator goes to ∞. Furthermore, BDSE has a normal limit distribution different from that of the global estimator which converges to a Chernoff limit, discussed in details below. In related work, [20] study M-estimation in non-standard cube-root problems of the type considered originally by [11] and show that the pooled-by-averaging estimator in a distributed computing framework has a different (in fact, normal) asymptotic distribution, as compared to the global M-estimator which converges to the unique maximizer of an appropriate mean 0 Gaussian process minus a quadratic drift.

Our goal in this paper is to propose *new estimators* under the DC framework in both the monotone function estimation problem as well as in certain versions of the M-estimation setting of [20] which do not suffer from the super-efficiency problem of the pooled-by-averaging estimators and which also recover the limiting properties of the corresponding global estimators. To this end, we first

provide some details of the problem considered in BDS and the results obtained therein, as well as those dealt with in [20], as they are crucial to understanding the goal and the approach of the current work.

Consider a sample of size N (very large) from the model $Y_i = \mu(X_i) + \epsilon_i$ which is distributed across m different servers, each server containing a subsample of size n, and $m = o(N)$. The function μ is known to be monotone and the X_i come from a density on $[0,1]$. The residual ϵ_i is assumed to satisfy $E(\epsilon_i|X_i) = 0$. Computing the global isotonic estimate at a point $t_0 \in (0,1)$ involves moving all the data to a central server and performing the isotonization on all N data-points on the central server. This can be time-consuming when N is really large. The construction of BDSE involves computing the isotonic estimate of μ, say $\widehat{\mu}_j$, on the j'th server, and then obtaining the average of these isotonic estimates. Hence, the pooled estimate at the point t_0 is given by: $\overline{\mu}(t_0) := m^{-1} \sum_{j=1}^{m} \widehat{\mu}_j(t_0)$. Computing BDSE at a particular point only requires transferring m numbers (from the m machines) to the central server, where $m = o(N)$.

One can compare the computational burden involved in the calculation of the global estimator to that for BDSE. For the global estimator, once all the data-points have been transferred to the central machine, sorting of the X_i's (resulting in an induced sorting of the Y_i's) can be accomplished typically in $O(N \log N)$ time. Post-sorting, one can implement isotonic regression via the PAVA algorithm [17] (Chapter 1) which takes $O(N)$ time. Thus, the total computational burden is $O(N \log N)$ computing time plus the transferring of N bivariate pairs to the central machine. On the other hand, for the pooled estimator, on each machine, the isotonic estimate based on the subsample stored in that machine takes $O(n \log n)$ computing time, leading to a total computing time of $O(mn \log n)$. At the central server, averaging takes $O(m)$ time. If $n \sim N^{\gamma}$ for some $0 < \gamma < 1$, this gives a total computing time of order $O(N \log N)$, and in addition, one transfers $m \sim N^{1-\gamma}$ scalars (the values $\widehat{\mu}_j(t_0)$ for $j = 1, 2, \ldots, m$) to the central machine. Thus, the pooled estimator is computationally less burdensome than the global estimator. Similar considerations apply to the computation of the global and pooled isotonic estimators of the inverse function μ^{-1}.

BDS showed that their pooled-by-averaging estimator (BDSE) of the inverse function has *dichotomous behavior*. We briefly revisit this important result. For convenience and the sake of completeness, we state these results essentially in their entirety.

Consider a nonincreasing and continuously differentiable function μ_0 on $[0,1]$ with $0 < c < |\mu_0^{'}(t)| < d < \infty$ for all $t \in [0,1]$. For an $x_0 \in (0,1)$, 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 $[0,1]$, coincide with μ_0 outside of $(x_0 - \epsilon_0, x_0 + \epsilon_0)$ for some (small) $\epsilon_0 > 0$, satisfy $0 < c < |\mu'(t)| < d < \infty$ for all $t \in [0,1]$, and such that $\mu^{-1}(a) \in (x_0 - \epsilon_0, x_0 + \epsilon_0)$ where $a = \mu_0(x_0)$. Now, consider N i.i.d. observations $\{(X_i, Y_i)\}_{i=1}^{N}$ from (X, Y) where $Y_i = \mu_0(X_i) + \epsilon_i$ and $X_i \sim$ Uniform$(0,1)$ is independent of $\epsilon_i \sim N(0, v^2)$. Then, the isotonic

estimate $\hat{\theta}_N$ of $\theta_0 := \mu_0^{-1}(a)$ (which is x_0) satisfies

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

as $N \to \infty$, where $G =_d \tilde{\kappa}\,\mathbb{Z}$, with $\mathbb{Z}$ following the Chernoff distribution, and $\tilde{\kappa} > 0$ being a constant. Writing $N = m \times n$, where m and n are as defined above, as $N \to \infty$, the BDSE of $\mu^{-1}(a)$, say $\overline{\theta}_m$ satisfies:

$$N^{1/3}(\overline{\theta}_m - \theta_0) \to_d m^{-1/6} H\,,$$

where H has the same variance as G but is distributed differently from G. Furthermore,

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

as $N \to \infty$. Hence, BDSE *outperforms* the global inverse isotonic regression estimator in terms of point wise MSE.

This phenomenon is *reversed* when one looks at the maximal MSEs of the two estimators over the class of models defined by $\mathcal{M}_0$, as described in Theorem 5.1 of BDS.

Theorem 1.1 (Theorem 5.1 of BDS)**.** *Let*

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

and

$$E_m := \liminf_{N\to\infty} \sup_{\mu\in\mathcal{M}_0} \mathbb{E}_\mu\left[N^{2/3}(\overline{\theta}_m - \mu^{-1}(a))^2\right]$$

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,

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

Therefore, from Theorem 1.1 it follows that the asymptotic maximal risk of BDSE diverges to ∞ at rate (at least) $m^{2/3}$. Thus, the better off we are with BDSE for a fixed function, the worse off we are in the uniform sense over the class of functions $\mathcal{M}_0$. Hence, unfortunately, while maintaining a computational burden that is better than the global estimator, BDSE has undesirable statistical properties as seen above.

As discussed above, similar phenomena arise in the cube-root M-estimation problems of the Kim and Pollard type [11], studied in [20]. In [11] the M-estimator is defined as the location of the maximum of an empirical process

$$\mathbb{P}_N(g,\theta) = \frac{1}{n}\sum_{i\leq N} g(X_i,\theta)$$

where $\theta \in \Theta \subset \mathbb{R}^d$, $X_1, \ldots, X_N$ are i.i.d. random variables and $g(\cdot, \theta)$ is the criterion function that is optimized. The maximizer $\hat{\theta}$, say, is estimating θ_0, the unique maximizer of $P\, g(X_1, \theta)$ where P is the common distribution of the X_i's. The cube-root convergence rate is an outcome of the so-called 'sharp-edge effect' – the fact that the g's are discontinuous in θ, coupled with a quadratic decline in $\|\theta - \theta_0\|^2$ of $P\, g(X_1, \theta)$ around θ_0. In the distributed computing setting of [20], the N observations are stored across m servers with the j'th server containing n_j observations, and the M-estimator $\hat{\theta}^{(j)}$ on the j'th server maximizes

$$\mathbb{P}_{n_j}^{(j)}\, g(\cdot, \theta) \equiv \frac{1}{n_j} \sum_{i=1}^{n_j} g(X_i^{(j)}, \theta)\,,$$

where $\{X_i^{(j)}\}_{i=1}^{n_j}$ are the data on the j'th server. The pooled estimator $\hat{\theta}_0 := \sum_{j=1}^{m} \omega_j \hat{\theta}^{(j)}$, where $\omega_j = n_j^{2/3} / \sum_k n_k^{2/3}$, reduces to the simple average when all subsamples have the same size. Theorem 2.1 of [20] provides the asymptotic distribution of the pooled estimator: it is seen that the estimator converges at rate $m^{-1/2}\, n^{-1/3}$ to a normal distribution, faster than the $N^{-1/3}$ rate of the global estimator. This parallels the results established in Sections 3 and 4 of BDS in the isotonic regression context.

As in BDS, [20] also encounter the super-efficiency phenomenon, which is discussed in Section B of their supplement for the location estimator (Section B.1) and the value-search estimator (Section B.2), two of the examples treated in their paper. They demonstrate in both problems that the maximal MSE of the pooled-by-averaging estimator over a collection of models in a neighborhood of a fixed model diverges to ∞ with N, while the maximal MSE of the global estimator remains bounded.

In both BDS and [20], super-efficiency results from computing the non-standard estimator at each local machine and then averaging these estimators at the central server. To avoid this undesirable phenomenon, the key idea is to *reverse* these steps, i.e., first average the data on each local server in an appropriate manner (which will typically depend on the structure and the dimension of the problem) to obtain *essentially sufficient* summary statistics which are then transferred to the central server. The summary statistics are now used to compute a non-standard 'pooled estimator' (via an adaptation of the M-estimation procedure used to solve for the global estimator) that replicates the properties of the global estimator and manages to avoid super-efficiency. The term *essentially sufficient* is used in the sense that these summary statistics are enough to compute an estimator that matches the performance of the global estimator. We will illustrate the idea in details for the isotonic regression problem studied in BDS and the location search problem considered by [20], but the prescription itself can be expected to work in a broader class of problems, subject to appropriate fine-tuning. Furthermore, for our analysis, we address a *broader scenario beyond i.i.d. data*. Since we are thinking of large N problems, with the data being stored separately in different servers, it is natural to allow heterogeneity across servers. Thus, while our N observations will be assumed to be independent, we will *no longer consider them to*

be identically distributed; rather, they will be assumed to come from a number of different (m) sub-populations with the data within each sub-population being i.i.d. The different sub-populations will be linked by a common parameter of interest. Furthermore, the N pairs will be scrambled across a number of different servers (say L), with the same server hosting data from different sub-populations, as well as data from the same sub-population potentially stored on multiple servers[1].

2. The isotonic regression problem

2.1. The new estimator for the regression function

Assume that we have m samples of respective sizes $n_1, \ldots, n_m$ and that for all $j = 1, \ldots, m$, the j-th sample is composed of i.i.d. pairs of real valued random variables (X_{ji}, Y_{ji}), $i = 1, \ldots, n_j$, such that $E(Y_{ji}|X_{ji}) = \mu(X_{ji})$ for all i, j and an unknown regression function μ defined on $[0, 1]$. We denote by F_{Xj} the common distribution function of the covariates X_{ji}, $i = 1, \ldots, n_j$ in the j-th sample. The data are stored on several servers numbered $1, \ldots, L$ for some integer $L \geq 1$. The allocation of data on the different servers is arbitrary in the sense that a sample can be spread on several servers, a server can host data from several different samples, and the number of stored observations can vary across the different servers. The number L of different servers can even grow as $N \to \infty$. The total sample size is $N = \sum_{j=1}^{m} n_j$.

For ease of exposition, when considering simultaneously all the samples, we relabel the observations from the m samples to obtain independent pairs (X_i, Y_i), $i = 1, \ldots, N$ such that $E(Y_i|X_i) = \mu(X_i)$, where the distribution function of X_i is one of $F_{X1}, \ldots, F_{Xm}$. Let K be a positive integer that grows to infinity as $N \to \infty$, and for all $k \in \{1, \ldots, K\}$, let $I_k = ((k-1)/K, k/K]$. Let S_ℓ denote the set of indices i, such that (X_i, Y_i) is stored in the ℓ'th server. Now, for each server ℓ ($1 \leq \ell \leq L$) record

$$T_{\ell k} = \sum_{i=1}^{N} Y_i \mathbb{1}_{i \in S_\ell} \mathbb{1}_{X_i \in I_k}$$

and

$$C_{\ell k} = \sum_{i=1}^{N} \mathbb{1}_{i \in S_\ell} \mathbb{1}_{X_i \in I_k},$$

for $k \in \{1, \ldots, K\}$. Next, for each ℓ, transfer $\{(T_{\ell k}, C_{\ell k})\}_{k=1}^{K}$ to a central server. Compute a regressogram estimate on the central server in the following manner:

[1] In BDS, the number of servers was designated by m, while in this paper we change notation and call it L. As we will see below, it is the number of different sub-populations that really enters into the properties of the pooled estimator in general and *not* the number of servers. When each sub-population has its own server, then obviously $L = m$.

for each $k \in \{1, \ldots, K\}$,

$$\begin{aligned} \overline{y}_k &= \frac{1}{\sum_{\ell=1}^{L} C_{\ell k}} \sum_{\ell=1}^{L} T_{\ell k} \\ &= \frac{1}{\sum_{i=1}^{N} \mathbb{1}_{X_i \in I_k}} \sum_{i=1}^{N} Y_i \mathbb{1}_{X_i \in I_k}. \end{aligned}$$

Our final estimator of $(\mu(\overline{x}_1), \ldots, \mu(\overline{x}_K))^T$, where $\overline{x}_k = k/K$, is

$$\widehat{y} = \arg \min_{h \in \mathbb{R}^K : h_1 \geq \cdots \geq h_K} \sum_{k=1}^{K} w_k (\overline{y}_k - h_k)^2 , \tag{2.1}$$

where

$$w_k = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{X_i \in I_k} = \frac{\sum_{\ell=1}^{L} C_{\ell k}}{N} .$$

The estimator of the regression fuction μ is obtained by piecewise-constant interpolation.

Note that the estimator does not depend on the way the observations were stored across different servers.

2.2. Computational considerations

Consider the computational burden for the new estimator. Assume, for now, that $K \sim N^{\zeta}$ for some $0 \leq \zeta < 1$. First, focus on the computational time it takes for calculating $(T_{\ell k}, C_{\ell k})$ for all ℓ and $1 \leq k \leq K$. For each X_i, one has to determine in which interval I_k it falls, and then assign the pair (X_i, Y_i) to the interval I_k. This can be accomplished in $O(\log N^{\zeta}) = O(\log N)$ time. Since there are N such points (scrambled across the different servers), the total time taken is $O(N \log N)$. Next, computing $(C_{\ell k}, T_{\ell k})$ for a fixed ℓ involves less than $2\, n_{\ell k}$ additions, where $n_{\ell k}$ is the number of (X_i, Y_i) pairs assigned to I_k on server ℓ. Hence, computing the vector $\{C_{\ell k}, T_{\ell k}\}_{1 \leq k \leq K}$ takes $O(\sum_k n_{\ell k})$ time. Summing up across the different ℓ's, we are looking at a total of $O(N \log N) \vee O(N)$ time, i.e. $O(N \log N)$ time.

After the pairs $\{T_{\ell k}, C_{\ell k}\}_{1 \leq k \leq K}$ have been transferred to the central server, computing the vector $\{(w_k, \overline{y}_k)\}_{1 \leq k \leq K}$ takes $O(LN^{\zeta})$ time, and the final isotonization step takes $O(N^{\zeta})$ time. Thus, the total computing time is $O(LN^{\zeta}) \vee O(N \log N)$ which is dominated by $O(N \log N)$ provided L (which could grow with N) and ζ are not too large. In addition to the total computing time, the burden also involves transferring about $2LK \sim LN^{\zeta}$ numbers between machines, which is larger than the amount of data transferred in the construction of BDSE. As shall be seen below, with K slightly larger than $N^{1/3}$ – say $K \sim N^{1/3+\eta_1}$ (η_1 small) – and m of a smaller order than $N^{1/3}$, the new estimator is able to recover the properties of the global estimator: hence, so long as the number of

machines is not too large – say $L = N^{1/3-\eta_2}$ – the total amount of data required to be transferred is of order $N^{2/3+\eta_1-\eta_2} = o(N^{2/3})$ when $\eta_2 > \eta_1$.

Note that the computation of the global isotonic estimator in this situation would require transferring all data points to the central server which is exactly $O(N)$ and the isotonic algorithm at the central server would take $O(N \log N)$ time. Note also that the minimum amount of data transferring needed for the new estimator above is of order K (this happens when the number of servers L is held fixed) and therefore of *larger order than* $N^{1/3}$. On the other hand, in the scenario of BDS, where $L = m$, the BDSE is constructed using m sub-samples where m is of order at most $N^{1/4}$: this corresponds to a data-transfer of *order at most* $N^{1/4}$ numbers to construct the super-efficient estimator at any given point. The additional amount of data that needs to be transferred to construct the new estimator can be viewed as the cost of alleviating the super-efficiency phenomenon exhibited by BDSE.

2.3. Characterization of the new estimators

It is a standard result in isotonic regression that the minimum in (2.1) is achieved at a unique vector $(\widehat{y}_1, \ldots, \widehat{y}_K)^T$. We give below a characterization of the minimizer. In the sequel, we consider the piecewise-constant left-continuous estimator $\widehat{\mu}_N$ that is constant on the intervals $[0, \overline{x}_1]$, and $(\overline{x}_{k-1}, \overline{x}_k]$ for all $k = 2, \ldots, K$, and such that

$$\widehat{\mu}_N(\overline{x}_k) = \widehat{y}_k$$

for all $k = 1, \ldots, K$. Let F_N be the empirical distribution function corresponding to $X_1, \ldots, X_N$

$$F_N(x) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{X_i \le x}, \ x \in \mathbb{R}, \tag{2.2}$$

and let Λ_N be the piecewise-constant right-continuous process on $[0, 1]$ that is constant on the intervals $[0, \overline{x}_1)$, and $[\overline{x}_{k-1}, \overline{x}_k)$ or all $k = 2, \ldots, K$ such that

$$\Lambda_N(\overline{x}_j) = \sum_{k=1}^{j} w_k \overline{y}_k = \frac{1}{N} \sum_{i=1}^{N} Y_i \mathbb{1}_{X_i \le \overline{x}_j}$$

for all $j = 1, \ldots, K$, and $\Lambda_N(0) = 0$. Then,

$$F_N(\overline{x}_j) = \sum_{k=1}^{j} w_k$$

and $\widehat{\mu}_N$ is the left-hand slope of the least concave majorant of the cumulative sum diagram defined by the set of points $\{(F_N(\overline{x}_k), \Lambda_N(\overline{x}_k)), k = 0, \ldots, K\}$ where $\overline{x}_0 = 0$. We define the corresponding inverse estimator as follows:

$$U_N(a) = \operatorname*{argmax}_{u \in \{\overline{x}_0, \ldots, \overline{x}_K\}} \{\Lambda_N(u) - aF_N(u)\} \tag{2.3}$$

where $\overline{x}_0 = 0$, argmax denotes the greatest location of the maximum, and where we 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. To see that $U_N = \widehat{\mu}_N^{-1}$, note that from the characterization above of $\widehat{\mu}_N$ as the slope of a least concave majorant, it follows that for all $a \in \mathbb{R}$ and $t \in (0,1]$, we have the equivalences

$$\begin{aligned}\widehat{\mu}_N(t) < a &\Leftrightarrow (\exists \overline{x}_i < t)\,(\forall \overline{x}_j \geq t) : \frac{\Lambda_N(\overline{x}_j) - \Lambda_N(\overline{x}_i)}{F_N(\overline{x}_j) - F_N(\overline{x}_i)} < a \\ &\Leftrightarrow (\exists \overline{x}_i < t)\,(\forall \overline{x}_j \geq t) : \Lambda_N(\overline{x}_j) - aF_N(\overline{x}_j) < \Lambda_N(\overline{x}_i) - aF_N(\overline{x}_i) \\ &\Leftrightarrow \operatorname*{argmax}_{u \in \{\overline{x}_0, \ldots, \overline{x}_K\}} \{\Lambda_N(u) - aF_N(u)\} < t\end{aligned}$$

whereas for $t = 0$, we have the equivalence

$$\widehat{\mu}_N(0) < a \Leftrightarrow \operatorname*{argmax}_{u \in \{\overline{x}_0, \ldots, \overline{x}_K\}} \{\Lambda_N(u) - aF_N(u)\} = 0.$$

We study below the asymptotic properties of $U_N(a)$ for arbitrary a and use these to deduce the asymptotic properties of $\widehat{\mu}_N(t)$ for a fixed $t \in (0,1)$ using the switch relation

$$\widehat{\mu}_N(t) \geq a \Longleftrightarrow t \leq U_N(a), \tag{2.4}$$

that holds for all $t \in (0,1]$ and $a \in \mathbb{R}$.

It will be useful to also record similar characterizations of the global estimator $\hat{\mu}_{N,G}$ of μ, for the sake of completeness. Recall that the global estimator is the isotonic estimator that we would compute if all the data $\{X_i, Y_i\}_{i=1}^N$ could have been brought over (or were already there) on a central server. Letting $\Lambda_{N,G}(t) = N^{-1} \sum_{i=1}^N Y_i \mathbb{1}_{X_i \leq t}$, for $a \in \mathbb{R}$, define

$$U_{N,G}(a) = \operatorname*{argmax}_{u \in [0,1]} \{\Lambda_{N,G}(u) - aF_N(u)\}. \tag{2.5}$$

Then $U_{N,G}(a) = \hat{\mu}_{N,G}^{-1}(a)$ and similar to the pooled estimator, we have the following characterization:

$$\widehat{\mu}_{N,G}(t) \geq a \Longleftrightarrow t \leq U_{N,G}(a), \tag{2.6}$$

that holds for all $t \in (0,1]$ and $a \in \mathbb{R}$.

3. Asymptotic properties of the new estimators

3.1. Notation and assumptions

In the sequel, we denote by g the generalized inverse of μ and by $\mathbb{E}^X$ the conditional expectation given $X_1, \ldots, X_N$. Being the inverse of μ, g is only defined

on the interval $[\mu(1), \mu(0)]$. In the sequel, we expand g to the whole real line by setting $g(a) = 0$ for all $a > \mu(0)$ and $g(a) = 1$ for all $a < \mu(1)$.

Furthermore, for all $x \geq 0$, $[x]$ denotes the integer part of x. We denote by F_X the mixing distribution function

$$F_X(x) = \sum_{j=1}^{m} \frac{n_j}{N} F_{Xj}(x). \tag{3.1}$$

Note that the function depends on N but for notational convenience, this is not made explicit in the notation.

To develop the asymptotic properties of the proposed estimator, we will impose some further conditions on the model. These are:

A1. Assume that F_X has a density function f_X on $[0, 1]$ that satisfies

$$C_1 < \inf_{t\in[0,1]} f_X(t) \leq \sup_{t\in[0,1]} f_X(t) \leq C_2 \tag{3.2}$$

for some positive numbers C_1 and C_2 that do not depend on N.

A2. With $\varepsilon_i = Y_i - \mu(X_i)$ for all $i = 1, \ldots, N$, assume that there exists $\sigma > 0$ such that $\mathbb{E}[\varepsilon_i^2 | X_i] \leq \sigma^2$ for all i, with probablity one.

A3. The regression function μ satisfies:

$$C_3 < \left| \frac{\mu(t) - \mu(x)}{t - x} \right| < C_4 \text{ for all } t \neq x \in [0, 1]\,, \tag{3.3}$$

for positive numbers C_3 and C_4.

A4. The number of bins K satisfies $K^{-1} = o(N^{-1/3})$ and there exists $\lambda \in (0, 1]$ that may depend on N and satisfies

$$\min_{1\leq j\leq m} \frac{n_j}{N} \geq \lambda > 0 \text{ and } \liminf_{N\to\infty} N^{1/3}\lambda(\log N)^{-3} = \infty. \tag{3.4}$$

Remarks on the assumptions: Assumption (A1) is fulfilled for instance if each $F_{Xj},\ j = 1, \ldots, m$ has a density function f_{Xj} such that $C_1 < f_{Xj}(x) < C_2$ for all $x \in [0, 1]$. Note that Assumption (A3) is weaker than differentiability, it implies that μ is both Lipschitz and so to speak *inverse Lipschitz*. It also implies that the inverse function g defined above is continuous. Assumption (A4) is critical to recovering the Chernoff-type asymptotics for the pooled estimator; that K grows faster than $N^{1/3}$ ensures that the data are averaged over bins of length smaller than $N^{-1/3}$, so that the isotonic algorithm operating on these averages at the central machine can still recover the $N^{-1/3}$ convergence rate. If K were to grow exactly at the rate $N^{1/3}$ or slower, the pooled estimator would no longer demonstrate Chernoff-type cube-root asymptotics. Furthermore, in (A4), we assume that the proportion n_j/N of observations from the j-th sample is at least of order $N^{-1/3}(\log N)^3$. This also plays a critical role in the subsequent analysis. Since,

$$1 = \sum_{j=1}^{m} \frac{n_j}{N} \geq m \min_{1\leq j\leq m} \frac{n_j}{N},$$

the conditions in (3.4) imply that the number m of different sub-samples cannot grow to fast: we must have $m \ll N^{1/3}(\log N)^{-3}$.

3.2. Uniformly bounded MSE property of the new estimators

The Inverse Problem: We first demonstrate that the new estimator in the inverse problem exhibits *uniformly bounded maximal risk (MSE)* over an appropriate class of models, as N grows to ∞. This is an analogue of the first result in Theorem 4.1 of BDS for the *global isotonic estimator of the inverse function*, though it is established here under weaker conditions. For this task, we denote by $\mathcal{F}_1$ the class of non-increasing functions μ on $[0,1]$ that satisfy (3.3) and $\sup_t |\mu(t)| \le C_5$, where $C_5 > 0$ is a positive number. The proof of the following theorem is given in Section 6.

Theorem 3.1. *Under assumptions (A1) through (A4), there exists $C > 0$ that depends only on $\sigma^2, C_1, C_2, C_3, C_4$ such that for all $a \in \mathbb{R}$,*

$$\limsup_{N\to\infty} \sup_{\mu\in\mathcal{F}_1} N^{2/3}\mathbb{E}_\mu(U_N(a) - \mu^{-1}(a))^2 \le C.$$

The Direct Problem: An analogue of the second result in Theorem 4.1 of BDS that demonstrates that the new estimator fixes the super-effciency phenomenon in the *direct problem* as well, i.e. $\hat{\mu}_N$ has bounded uniform MSE as $N \to \infty$ over the class $\mathcal{F}_1$, can also be established. As it involves some additional technical fine-tuning we relegate its proof to the Appendix.

Theorem 3.2. *Fix $\delta \in (0, 1/2)$. Then, there exists $C > 0$ that depends only on $\sigma, p, C_1, C_2, C_3, C_4, C_5, \delta$ such that for all $t \in [\delta, 1-\delta]$*

$$\limsup_{N\to\infty} \sup_{\mu\in\mathcal{F}_1} N^{2/3}\mathbb{E}_\mu(\widehat{\mu}_N(t) - \mu(t))^2 \le C.$$

We show next that under a fixed μ, the new estimator recovers the asymptotic distribution of the global estimator with the same convergence rate.

3.3. Asymptotic distributions

To establish asymptotic distributions for our new estimators, we make *additional assumptions* in the case that the number m of different samples goes to infinity, and we clarify the asymptotic setting further.

When considering the case where m is allowed to grow to infinity as $N \to \infty$, we assume that there is a sequence of unknown distinct distributions $\{P_j\}_{j\ge1}$ such that our set of observations is part of an infinite sequence of pairs $\{(X_i, Y_i)\}_{i\ge1}$, where for all i the distribution of (X_i, Y_i) takes the form P_j for some $j \ge 1$. Hence, $m = m_N$ is the number of different distributions that appear across the first N observations $(X_1, Y_1), \ldots, (X_N, Y_N)$. To fix ideas, possibly rearranging the probabilities in the sequence $\{P_j\}_{j\ge1}$, we assume without loss of

generality in the sequel that for all N, the $m = m_N$ distributions that appear across the first N observations are $P_1, \ldots, P_m$. Note that the setting does not exclude that $m_N = 1$ for all N, i.e. that all observations are drawn from the same distribution P_1. In the case where $m_N > 1$ for sufficiently large N, it is not excluded that m_N remains bounded. In the sequel, for all $j \geq 1$, we denote by σ_j the function such that

$$\sigma_j^2(u) = \mathbb{E}[(Y - \mu(X))^2 | X = u]$$

for all $u \in [0,1]$ and by f_j the density function of X, which is assumed to exist, where (X, Y) has distribution P_j. Then, the distribution function F_X in (3.1) has a density function f_X on $[0,1]$ given by

$$f_X(u) = \sum_{j=1}^{m} \frac{n_j}{N} f_j(u) \tag{3.5}$$

for all $u \in [0,1]$.

We next make the following technical assumptions.

$\tilde{A}_0$. The functions $\{f_j\}$ are uniformly bounded in j on the interval $[0,1]$.

$\tilde{A}_1$. Let

$$\omega(\delta) = \sup_{j \geq 1} \max\{ \sup_{|u-v| \leq \delta} |\sigma_j^2(u) - \sigma_j^2(v)|, \sup_{|u-v| \leq \delta} |f_j(u) - f_j(v)| \}$$

for all $\delta \geq 0$. Then, $\omega(\delta) \to 0$ as $\delta \to 0$.

$\tilde{A}_2$. The density function f_X converges pointwise [and hence, uniformly] on $[0,1]$ as $N \to \infty$ to a continuous function f_∞ that is bounded away from zero. This implies that (3.2) holds for some positive numbers C_1, C_2 that do not depend on N, provided that N is sufficiently large.

$\tilde{A}_3$. The function σ_X^2 defined by

$$\sigma_X^2(u) := \sum_{j=1}^{m} \frac{n_j}{N} \sigma_j^2(u) f_j(u)$$

for all $u \in [0,1]$ converges pointwise [and hence, uniformly] to a continuous function σ_∞^2, bounded away from 0, as $N \to \infty$.

$\tilde{A}_4$. With $\varepsilon_i := Y_i - \mu(X_i)$ for all $i = 1, \ldots, N$, there exists $\sigma > 0$ such that $\mathbb{E}[|\varepsilon_i|^p | X_i = t] \leq \sigma^p$ for all i, t and some $p > 2$.

$\tilde{A}_5$. The function μ is decreasing and has a continuous first derivative on $[0,1]$ such that $\inf_{u \in [0,1]} |\mu'(u)| > 0$

For notational convenience, we do not make it explicit in the notation that F_X, f_X, σ_X, m may depend on N.

Remark: The pointwise convergence of f_X to f_∞ implies uniform convergence because by assumption $\tilde{A}_1$, the class of functions $\{f_j\}$ is uniformly equicontinuous, which then implies that the class $\{f_X\}$ is also uniformly equicontinuous.

Also, the pointwise convergence of σ_X^2 to σ_∞^2 guarantees uniform convergence, because the class of functions $\{\sigma_X^2\}$ is uniformly equicontinuous: this follows from the uniform boundedness of the class $\{f_j\}$ assumed in $\tilde{A}_0$, the uniform boundedness of $\{\sigma_j^2\}$, which is a consequence of $\tilde{A}_4$, and the uniform equicontinuity of the classes $\{f_j\}$ and $\{\sigma_j^2\}$ assumed in $\tilde{A}_1$.

Theorem 3.3. *With $t \in (0,1)$ fixed, and $a = \mu(t) + N^{-1/3}x$ for some fixed $x \in \mathbb{R}$, under Assumptions $\tilde{A}_1$ through $\tilde{A}_4$ and A4, we have*

$$N^{1/3}(U_N(a) - g(a)) \to_d \left(\frac{2\sigma_\infty(t)}{|\mu'(t)|f_\infty(t)}\right)^{2/3} \mathbb{Z} \text{ as } N \to \infty,$$

where $\mathbb{Z} := \text{argmax}_{u\in\mathbb{R}}\{W(u) - u^2\}$, W being a standard two-sided Brownian motion starting at 0, has the so-called Chernoff's distribution.

An interesting feature of the estimator U_N is that its asymptotic behavior does not depend on the way the N data are allocated on the different servers. The direct estimator $\hat{\mu}_N$ shares this feature, as is shown in the next result.

Theorem 3.4. *Under the same assumptions as in Theorem 3.3, with $t \in (0,1)$ fixed, we have*

$$N^{1/3}(\widehat{\mu}_N(t) - \mu(t)) \to_d \left(\frac{4\sigma_\infty^2(t)|\mu'(t)|}{f_\infty^2(t)}\right)^{1/3} \mathbb{Z} \text{ as } N \to \infty,$$

where $\mathbb{Z}$ is as defined in Theorem 3.3.

Remark: The estimators $\widehat{\mu}_N(t)$ and $U_N(a)$ have the *same asymptotic distributions* (when centered around their respective estimands and scaled by the factor $N^{1/3}$) as the corresponding global isotonic estimators, $\hat{\mu}_{N,G}$ and $U_{N,G}$ defined in (2.6) and (2.5) respectively. In other words, the asymptotic distributions of the estimators $N^{1/3}(U_{N,G} - g(a))$ and $N^{1/3}(\hat{\mu}_{N,G}(t) - \mu(t))$ are those arising in Theorems 3.3 and 3.4 respectively. The limit distributions of the global estimators can be established by the same set of techniques as used in the proofs of Theorems 3.3 and 3.4. Thus, the new estimators proposed in this paper not only circumvent the super-efficiency phenomenon but recover the asymptotic properties of their corresponding global versions. We note that the global isotonic estimators $\widehat{\mu}_{N,G}(t)$ and $U_{N,G}(a)$ also possess the *uniformly bounded maximal MSE* property for their respective estimands, i.e. exact analogues of the results in Theorems 3.2 and 3.1 hold for $N^{1/3}(U_{N,G} - g(a))$ and $N^{1/3}(\hat{\mu}_{N,G}(t) - \mu(t))$ respectively, and can be established by similar techniques as used in the proofs of these two theorems.

Remark: The setting of the theorems in this section with a growing sequence of sub-populations such that conditions $\tilde{A}_1$ through $\tilde{A}_5$ hold is not difficult to satisfy. Consider, for example, $m = \lfloor N^{1/4} \rfloor$ and P_j has density $f_j(u) = (1 - \epsilon_j)f_0(u) + \epsilon_j f_1(u)$ where f_0 and f_1 are Lipschitz continuous densities bounded away from 0 and ∞ on $[0,1]$, $0 < \epsilon_j < 1$ for all j, the sequence $\{\epsilon_j\}$ is decreasing

to 0 and $\sum_{j=1}^{m} \epsilon_j = o(m)$, which is easy to arrange. Let the distribution of the X_i's be P_1 for $i = 1, 2, \ldots, \lfloor N/m \rfloor$, P_2 for $\lfloor N/m \rfloor + 1 \leq i \leq 2\lfloor N/m \rfloor$, ..., and P_m for $(m-1)\lfloor N/m \rfloor \leq i \leq N$. For each i, the regression model is $Y = \mu(X_i) + \epsilon_i$ where the ϵ_i's are i.i.d. $N(0, \sigma^2)$ (say) and independent of the X_i's, which are also mutually independent, and μ satisfies all the desired conditions in this manuscript, in particular $\tilde{A}_5$. Then, it is easy to check that all the five conditions at the beginning of this section hold, with $f_\infty = f_0$ and $\sigma^2_\infty(u) = \sigma^2 f_\infty(u)$.

The proof of Theorem 3.3 is in the Appendix. The proof of Theorem 3.4 follows.

Proof of Theorem 3.4. It follows from the switch relation (2.4) that for all fixed $t \in (0,1)$, with $a = \mu(t) + N^{-1/3}x$ we have

$$\begin{aligned} \mathbb{P}\left(N^{1/3}(\widehat{\mu}_N(t) - \mu(t)) < x\right) &= \mathbb{P}\left(\widehat{\mu}_N(t) < \mu(t) + N^{-1/3}x\right) \\ &= \mathbb{P}\left(t > U_N(a)\right) \\ &= \mathbb{P}\left(N^{1/3}(U_N(a) - g(a)) < N^{1/3}(t - g(a))\right). \end{aligned}$$

Now, $N^{1/3}(t - g(a)) = xg'(\mu(t)) + o(1) = x|\mu'(t)|^{-1} + o(1)$, so it follows from Theorem 3.3 that

$$\lim_{N\to\infty} \mathbb{P}\left(N^{1/3}(\widehat{\mu}_N(t) - \mu(t)) < x\right) = \mathbb{P}\left(\left(\frac{2\sigma_\infty(t)}{|\mu'(t)|f_X(t)}\right)^{2/3} \mathbb{Z} < \frac{x}{|\mu'(t)|}\right),$$

using that the Chernoff distribution $\mathbb{Z}$ is continuous (see e.g. [9]). □

4. The location parameter problem

The location parameter problem is one of the examples studied by [20] (Section 3.1) and falls in the genre of general cube-root M-estimation problems introduced by [11]. As discussed in Section 1, in the framework of [11], a generic estimator maximizes an empirical process $\mathbb{P}_N(g, \theta) = \frac{1}{n}\sum_{i\leq N} g(\xi_i, \theta)$ over $\theta \in \Theta \subset \Theta$ with $\Theta \subset \mathbb{R}^d$, the ξ_i's being i.i.d. random variables. Consider the case $d = 1$, $\Theta = [0,1]$ and assume that the ξ_i's assume values in $[0,1]$. One particular recipe which (or embellishments of which) works in a variety of cases, e.g. the location parameter problem, where the global estimator is obtained by searching over the values of the ξ_i's, is to define $\overline{x}_k = k/K$ for $k = 1, \ldots, K$, compute $\sum_{i\leq N} g(\xi_i, \overline{x}_k)\mathbb{1}(i \in S_\ell)$ on each server $l \in \{1, \ldots, L\}$[2], transfer each summary to the central server, and sum up to obtain $\sum_{i\leq N} g(\xi_i, \overline{x}_k)$ for each k. The final estimator is computed as the argmax of $\sum_{i\leq N} g(\xi_i, \theta)$ over $\theta \in \{\overline{x}_1, \ldots, \overline{x}_K\}$. For this 'pooled estimator' to recover the properties of the global estimator, we would expect that $N^{1/3} = o(K)$, as in the isotonic regression problem. We present the detailed analysis below for location estimation.

[2] S_l is the set of indices i such that the corresponding ξ_i are on the l'th server.

4.1. The set-up, the estimator and assumptions

Assume that we have m samples of respective sizes $n_1, \ldots, n_m$ and that for all $j = 1, \ldots, m$, the j-th sample is composed of i.i.d. random variables X_{ji}, $i = 1, \ldots, n_j$, with common distribution P_j, such that for a fixed bandwidth r_0, there exists a unique θ_0 such that both $\theta_0 - r_0$ and $\theta_0 + r_0$ are in $(0, 1)$ and

$$\theta_0 = \arg\max_{\theta} P_j([\theta - r_0, \theta + r_0]).$$

One special case of the above is the situation that each P_j is unimodal with a common mode (across j). We denote by F_{Xj} the common distribution function of the variables X_{ji}, $i = 1, \ldots, n_j$ in the j-th sample. The data are stored on several servers numbered $1, \ldots, L$ for some integer $L \geq 1$, and we allow the possibility that L grows with $N \equiv \sum_{j=1}^{m} n_j$, the total sample size. The allocation of data on the different servers is arbitrary in the sense that a sample can be spread on several servers, a server can host data from several different samples, and the number of stored observations can vary across the different servers.

For ease of exposition, when considering simultaneously all the samples, we relabel the observations from the m samples to obtain independent variables X_i, $i = 1, \ldots, N$ whose the distribution function is one of $F_{X1}, \ldots, F_{Xm}$. Let K be a positive integer that grows to infinity as $N \to \infty$, and for all $k \in \{1, \ldots, K\}$, let $I_k = ((k-1)/K, k/K]$. Let S_ℓ denote the set of indices i, such that X_i is stored in the ℓ'th server. Now, for each server ℓ ($1 \leq \ell \leq L$) record

$$T_{\ell k} = \sum_{i=1}^{N} \mathbb{1}_{i \in S_\ell} \mathbb{1}_{X_i \in I_k}$$

for $k \in \{1, \ldots, K\}$. Next, for each ℓ, transfer $\{T_{\ell k}\}_{k=1}^{K}$ to a central server. Compute an empirical distribution function on the central server in the following manner: Λ_N is the piecewise-constant right-continuous process on $[0, 1]$ that is constant on the intervals $[\overline{x}_{k-1}, \overline{x}_k)$ or all $k = 1, \ldots, K$ where $\overline{x}_k = k/K$, such that

$$\Lambda_N(\overline{x}_j) = \frac{1}{N} \sum_{k=1}^{j} \sum_{\ell=1}^{L} T_{\ell k} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{X_i \leq \overline{x}_j}$$

for all $j = 1, \ldots, K$, and $\Lambda_N(0) = 0$. We define the estimator θ_N of θ_0 as follows:

$$\theta_N = \arg\max_{\theta} \{\Lambda_N(\theta + r_{0K}) - \Lambda_N(\theta - r_{0K})\}$$

where $r_{0K} = [Kr_0]/K$, argmax denotes the greatest location of the maximum and where the maximum is taken over $\{\bar{x}_0, \ldots, \bar{x}_K\} \cap [r_{0K}, 1 - r_{r0K}]$. Note that both $\theta_N - r_{0K}$ and $\theta_N + r_{0K}$ belong to the set $\{\bar{x}_0, \ldots, \bar{x}_K\}$.

In the sequel, we use the same notation F_X as in (3.1) and we make the following assumption:

S. The number of bins K satisfies $K^{-1} = o(N^{-1/3})$ and there exists $\lambda \in (0, 1]$ that may depend on N and satisfies

$$\min_{1 \leq j \leq m} \frac{n_j}{N} \geq \lambda > 0 \text{ and } \liminf_{N \to \infty} N^{1/3} \lambda (\log N)^{-3} = \infty. \tag{4.1}$$

4.2. Theoretical properties of the pooled estimator

We demonstrate below that the new estimator in the location parameter problem exhibits *uniformly bounded maximal risk (MSE)* over an appropriate class of models, as N grows to ∞. An asymptotic distributional result under some further assumptions along the lines of the results in Section 3.3 can also be established, but is skipped.

For fixed positive numbers $C_1, C_2, C_3, a, \epsilon, \delta$ that do not depend on N, we denote by $\mathcal{F}_1$ the class of functions f_X on $[0,1]$ that satisfy

$$f_X = \sum_{j=1}^{m} \frac{n_j}{N} f_{Xj}$$

where

$$C_1 < \inf_{t\in[0,1]} f_{Xj}(t) \leq \sup_{t\in[0,1]} f_{Xj}(t) \leq C_2 \qquad \text{for all } j\,, \tag{4.2}$$

f_X is differentiable in neighborhoods of $\theta_0 - r_0$ and $\theta_0 + r_0$ with derivative that satisfies

$$f'_X(\theta_0 - r_0) - f'_X(\theta_0 + r_0) > \epsilon, \quad \sup_{u\in(0,1)} |f'_X(u)| \leq C_3 \tag{4.3}$$

and for $u_0 \in \{\theta_0 - r_0, \theta_0 + r_0\}$,

$$\sup_{|u-u_0|\leq a} |f'_X(u) - f'_X(u_0)| \leq \epsilon/3, \tag{4.4}$$

and for the primitive F_X of f_X we have

$$\sup_{|u-\theta_0|\geq a/2} \{F_X(u + r_0) - F_X(\theta_0 + r_0) - F_X(u - r_0) + F_X(\theta_0 - r_0)\} < -\delta. \tag{4.5}$$

It follows from the definition of θ_0 and F_X that θ_0 is the unique location of the maximum

$$\theta_0 = \arg\max_{\theta} \{F_X(\theta + r_0) - F_X(\theta - r_0)\}. \tag{4.6}$$

Hence, the supremum in (4.5) is stricty negative for all $a > 0$, and we consider a class $\mathcal{F}_1$ of functions where it is uniformly negative. The first condition on (4.3) is satisfied with some (small) ϵ for instance if f_X is increasing on $[0, \theta_0]$ and decreasing on $[\theta_0, 1]$, whereas (4.4) holds if f'_X is continuous in neighborhoods of both $\theta_0 - r_0$ and $\theta_0 + r_0$, and a is chosen sufficiently small.

We denote by $\mathbb{E}_f$ the underlying expectation when the distributions of the observations are such that the true density f_X of F_X defined in (3.1) is equal to f.

Theorem 4.1 (Theorem 3.1 equivalent)**.** *Under the assumptions made in this section,*

$$\limsup_{N\to\infty} \sup_{f\in\mathcal{F}_1} N^{2/3} \mathbb{E}_f(\theta_N - \theta_0)^2 < \infty.$$

5. Discussion

We have proposed new estimators for distributed computing in the isotonic regression problem and a prototypical cube-root estimation problem of the genre considered in [11] that preserve the convergence rates of the corresponding global estimators and do not suffer from the super-efficiency phenomenon. The key change from the BDS procedure or the procedure in [20] lies in smoothing the data on local servers followed by solving a non-linear optimization problem on the central server. In the isotonic setting, this is referred to as an 'SI' (smoothing-isotonization) procedure. We note here that such 'SI' procedures and their converse ('IS') procedures have been studied in monotone function problems, though *not in distributed computing environments* and *not under the heterogeneity setting of our paper*. See, for example, [16], [14], [22], [1] and [8]. An interesting topic for future work is to understand distributed computing and inference for non-standard problems in higher dimensions, e.g. the maximum score estimator treated in both [11] and [20] where the parameter is a p-dimensional vector ($p > 1$). For example, the partitioning into bins strategy used above that works well for 1-dimensional problems has a downside in larger dimensions, since the bins become hyper-cubes whose number increases quickly with p. This entails increased levels of communication among the different machines that an effective distributed computing strategy would seek to avoid.

Reverting to the isotonic regression problem, the ideas in this paper also have connections to other work in the monotone function literature that are worth mentioning. [25] study isotonic estimation of a decreasing density with histogram-type data based on i.i.d. data under a once differentiable assumption on the density. The domain of the density is split into bins, and the counts in each bin are available. When the number of bins grows at a rate faster than $n^{1/3}$, Theorem 4.6 of this paper shows that the isotonic estimate based on binned data recovers the Chernoff-type asymptotic distribution of the classical Grenander estimator. A similar phenomenon transpires in our problem. The $(C_{\ell k}, T_{\ell k})$ pair records the number of observations in the bin I_k and the sum of the responses in that bin respectively, for the ℓ'th server. Once these are transferred to the central server, we sum across ℓ to find the total number of observations in I_k and the sum of the responses corresponding to all those observations and construct our isotonic estimator using these statistics. In our problem, K grows faster than $N^{1/3}$ and we obtain a Chernoff limit for the pooled estimator.

This naturally raises the question as to how the number of bins K for the smoothing step on the local servers would influence the distribution of the estimators developed in this paper. When $N^{1/3} = o(K)$, the grid is sufficiently dense and the corresponding bins sufficiently small, so that our isotonized regressogram estimator recovers the asymptotics of the classical, i.e. global isotonic regression estimator, but this will no longer be the case when $K \sim N^{1/3}$ or $K = o(N^{1/3})$. When $K \sim N^{1/3}$, the results of [25] (Theorem 3.3 and Corollary 4.4) and [21] (Theorem 3.7) who study monotone function estimation with covariates supported on a grid indicate that the limit distribution of the isotonized regressogram estimator at a point will neither be normal, nor will it be given

by Chernoff's distribution. When $K = o(N^{1/3})$, the grid is sparse enough and therefore, the regressogram estimates are ordered with probability increasing to one, so that the isotonized regressogram estimator agrees with the original estimator with increasing probability, and the results in [25] (Theorem 4.1) and [21] (Theorem 3.1) suggest an asymptotic normal distribution for our proposed estimator. We do not go into a full investigation of the details of these asymptotics in the distributed setting, since this is not relevant to the goal of the current work: produce a pooled estimator whose properties mimic the global estimator.

Some limited simulation results illustrating the role of K are presented in the Appendix. As noted above, for the pooled estimator to recover the properties of the global isotonic estimator, we need K to be of larger order than $N^{1/3}$, but to keep data-transfer costs low we would also like K to be not much larger. Since issues with distributed computing are only important for substantially large N, we investigated how our proposed estimator behaves in terms of K when N is in the order of millions or larger. It turns out that even a logarithmic adjustment, i.e. $K \sim N^{1/3} \log N$ performs very well: the resolution of the bins is good enough that the pooled estimator replicates the behavior of the global estimator to a high level of precision. In sum, the choice of K does not appear to be a critical issue in a really 'big data' setting. This is fortunate, as a heavy-duty tuning algorithm to determine K would enhance computational costs which one is trying to avoid in the first place. We also noted that changing to $K \sim N^{1/3}$ induces significant bias in our estimators (which is compatible with our observations in the previous paragraph).

As far as inference on the parameters of interest is concerned, the limit distributions, especially in the heterogeneous data setting contain several nuisance parameters which need to be estimated. Specifically the estimation of μ' in the isotonic regression problem is known to be difficult. One possibility in the isotonic regression problem is to use the likelihood ratio test for testing $H_0 : \mu(t_0) = \theta_0$ using the data at the central server, along the lines of the ideas developed in [5] and [4]. We believe that at least in the homogeneous setting, i.e. when the data across the different servers are i.i.d., this likelihood ratio statistic will be asymptotically pivotal. It is possible that pivotality also holds under the general heterogeneous framework of this paper, but this would require further investigation. A comprehensive treatment of effective inference strategies would be an exciting topic for future research. We note that likelihood ratio statistics in heterogeneous massive data settings, albeit in a different genre of problems have been studied elsewhere in the literature, see e.g. [13].

We believe that similar estimators can be proposed for distributed convex regression. For convex regression, a BDS type estimator is expected to fail completely, since the global convex least squares estimator is itself asymptotically biased, as suggested by the extensive simulation experiments in [2]. However, a convexified regressogram estimator in the spirit of the one considered in this paper, ought to be able to recover the properties of the global convex LS estimator provided K is selected appropriately: we conjecture that in the convex case K should be taken to be $N^{1/5} = o(K)$. This could provide a possible avenue for future research.

6. Proof of Theorem 3.1

The proof of the above theorem relies on a number of preliminary results which are presented, next. In the remainder of this section, we assume that assumptions (A1) to (A4) are always satisfied (though some results may require only a subset of these assumptions). Additional assumptions will be imposed when required.

Lemma 6.1. *Let $\theta > 0$ be arbitrary. Then, there exist (i) a number $c > 0$ that depends only on C_1, C_3, (ii) an integer $N_0 > 0$ that depends only on $C_1, C_2, C_3, C_4, \theta$ and (iii) an event $\mathcal{E}_N$ that depends only on C_2, such that for all $N \geq N_0$, we have $\mathbb{P}(\mathcal{E}_N) \geq 1 - N^{-\theta}$ and on $\mathcal{E}_N$,*

$$E^X \Lambda_N(u) - E^X \Lambda_N\left(\frac{[Kg(a)]}{K}\right) - a\left(F_N(u) - F_N\left(\frac{[Kg(a)]}{K}\right)\right) \leq -c(u - g(a))^2$$

for all $a \in \mathbb{R}$ and all $u \in \{\overline{x}_0, \ldots, \overline{x}_K\}$ such that $|u - g(a)| \geq N^{-1/3}$.

The proof of this lemma is long and technical and is available in the Appendix. The next result gives a polynomial tail bound on the estimation error $U_N(a) - g(a)$ over a high-probability set that is eventually used to bound the MSE.

Lemma 6.2. *With $\mathcal{E}_N$ and N_0 taken from Lemma 6.1, there exists $C > 0$ that depends only on $\sigma^2, C_1, C_2, C_3, C_4$ such that for all $a \in \mathbb{R}$ and $x > 0$,*

$$\mathbb{P}\left(|U_N(a) - g(a)| \geq x, \mathcal{E}_N\right) \leq \frac{C}{Nx^3} \tag{6.1}$$

for all $N \geq N_0$.

Proof of Lemma 6.2. The inequality in the lemma is obvious for $x \in (0, N^{-1/3})$ since for such x's, it suffices to choose $C \geq 1$ so that the right hand side is larger than one. Hence, in the sequel we consider $x \geq N^{-1/3}$. For all $a \in \mathbb{R}$ and all $u \in \{\overline{x}_0, \ldots, \overline{x}_K\}$ such that $|u - g(a)| \geq x$, define $e(a, u)$ as in (A.21) and $M_N(u) = \Lambda_N(u) - E^X(\Lambda_N(u))$. The characterization in (2.3) proves that the event $\{U_N(a) - g(a) \geq x\}$ is included in the event

$$\begin{aligned}\left\{\max_{u \in \{\overline{x}_0, \ldots, \overline{x}_K\},\ u - g(a) \geq x} \{\Lambda_N(u) - aF_N(u)\} \geq \Lambda_N\left(\frac{[Kg(a)]}{K}\right) - aF_N\left(\frac{[Kg(a)]}{K}\right)\right\} \\ = \left\{\max_{u \in \{\overline{x}_0, \ldots, \overline{x}_K\},\ u - g(a) \geq x} \left\{M_N(u) - M_N\left(\frac{[Kg(a)]}{K}\right) + e(a, u)\right\} \geq 0\right\}.\end{aligned}$$

Since $x \geq N^{-1/3}$, combining this with Lemma 6.1 shows that there exists $c > 0$ that depends only on C_1, C_3 such that with $\mathcal{X} = \{\overline{x}_0, \ldots, \overline{x}_K\}$,

$$\begin{aligned}&\mathbb{P}\left(U_N(a) - g(a) \geq x, \mathcal{E}_N\right) \\ &\quad \leq \mathbb{P}\left(\max_{u \in \mathcal{X},\ u - g(a) \geq x} \left\{M_N(u) - M_N\left(\frac{[Kg(a)]}{K}\right) - c(u - g(a))^2\right\} \geq 0\right)\end{aligned}$$

for $N \geq N_0$. The above probablity is less than or equal to

$$\sum_{k\geq 0} \mathbb{P}\left(\max_{u\in\mathcal{X},\ u-g(a)\in[x2^k,x2^{k+1}]} \left\{ M_N(u) - M_N\left(\frac{[Kg(a)]}{K}\right) - c(u-g(a))^2 \right\} \geq 0 \right)$$
$$\leq \sum_{k\geq 0} \mathbb{P}\left(\max_{u\in\mathcal{X},\ u-g(a)\in[0,x2^{k+1}]} \left\{ M_N(u) - M_N\left(\frac{[Kg(a)]}{K}\right) \right\} \geq c(x2^k)^2 \right). \tag{6.2}$$

Let $\mathbb{P}^X$ denote the conditional probability given $X_1, \ldots, X_N$. By definition, for all $u \in \{\overline{x}_0, \ldots, \overline{x}_K\}$ we have

$$M_N(u) = \frac{1}{N}\sum_{i=1}^{N} \varepsilon_i \mathbb{1}_{X_i \leq u} \tag{6.3}$$

where $\varepsilon_i = Y_i - \mu(X_i)$. The process M_n can be extended to all $u \in \mathbb{R}$ using the same definition as above. Then, M_N is a centered martingale under $\mathbb{P}^X$ that satisfies

$$\mathbb{E}^X \left(M_N(u) - M_N(v)\right)^2 = \frac{1}{N^2}\sum_{i=1}^{N} \mathbb{E}^X(\varepsilon_i^2)\mathbb{1}_{u<X_i\leq v} \leq \frac{\sigma^2}{N}(F_N(u) - F_N(v)) \tag{6.4}$$

for all $u \leq v$, using that $\mathbb{E}^X(\varepsilon_i^2) \leq \sigma^2$ for all i by assumption. Hence, it follows from the Doob inequality that for all $k \geq 0$,

$$\mathbb{P}^X \left(\max_{u\in\{\overline{x}_0,\ldots,\overline{x}_K\},\ u-g(a)\in[0,x2^{k+1}]} \left\{ M_N(u) - M_N\left(\frac{[Kg(a)]}{K}\right) \right\} \geq c(x2^k)^2 \right)$$
$$\leq \sigma^2 \frac{F_N\left(g(a) + x2^{k+1}\right) - F_N\left(\frac{[Kg(a)]}{K}\right)}{c^2 N (x2^k)^4}.$$

Taking the expectation on both sides of the preceding inequality yields for large enough N that

$$\mathbb{P} \left(\max_{u\in\{\overline{x}_0,\ldots,\overline{x}_K\},\ u-g(a)\in[0,x2^{k+1}]} \left\{ M_N(u) - M_N\left(\frac{[Kg(a)]}{K}\right) \right\} \geq c(x2^k)^2 \right)$$
$$\leq \frac{\sigma^2 \left\{ F_X\left(g(a) + x2^{k+1}\right) - F_X\left(\frac{[Kg(a)]}{K}\right) \right\}}{c^2 N (x2^k)^4}$$
$$\leq \frac{\sigma^2 C_2 (x2^{k+1} + K^{-1})}{c^2 N (x2^k)^4}$$
$$\leq \frac{2\sigma^2 C_2 x 2^{k+1}}{c^2 N (x2^k)^4},$$

where C_2 is taken from (3.2). For the penultimate inequality, we used that $x2^{k+1} \geq N^{-1/3}$ for all k whereas $K^{-1} = o(N^{-1/3})$, implying that $K^{-1} \leq x2^{k+1}$ for all k provided that N is sufficiently large. Putting the previous inequality in

(6.2) we obtain that for sufficiently large N,

$$\mathbb{P}\left(U_N(a) - g(a) \geq x, \mathcal{E}_N\right) \leq \sum_{k\geq 0} \frac{4\sigma^2 C_2}{c^2 N (x2^k)^3}.$$

Since $C := \sum_{k\geq 0} 2^{-3k}$ is finite, we conclude that

$$\mathbb{P}\left(U_N(a) - g(a) \geq x, \mathcal{E}_N\right) \leq \frac{4\sigma^2 C_2 C}{c^2 N x^3}.$$

Similar arguments show that

$$\mathbb{P}\left(g(a) - U_N(a) \geq x, \mathcal{E}_N\right) \leq \frac{4\sigma^2 C_2 C}{c^2 N x^3},$$

and therefore,

$$\mathbb{P}\left(|g(a) - U_N(a)| \geq x, \mathcal{E}_N\right) \leq \frac{8\sigma^2 C_2 C}{c^2 N x^3}.$$

The lemma follows. □

We are now ready to prove the theorem.

Proof of Theorem 3.1. Fix $\mu \in \mathcal{F}_1$ arbitrarily. Since both U_N and μ^{-1} take values in $[0,1]$, we have $|U_N(a) - \mu^{-1}(a)| \leq 1$ for all a and therefore, with $\overline{\mathcal{E}}_N$ the complementary event to $\mathcal{E}_N$ taken from Lemma 6.1, where we set $\theta = 2/3$, we have

$$\begin{aligned} \mathbb{E}_\mu\left(|U_N(a) - \mu^{-1}(a)|^2 \mathbb{1}_{\overline{\mathcal{E}}_N}\right) &\leq \mathbb{P}_\mu\left(\overline{\mathcal{E}}_N\right) \\ &\leq N^{-2/3} \end{aligned} \tag{6.5}$$

for N sufficiently large. On the other hand, it follows from the Fubini theorem that

$$\begin{aligned} \mathbb{E}_\mu\left(|U_N(a) - \mu^{-1}(a)|^2 \mathbb{1}_{\mathcal{E}_N}\right) &= \int_0^\infty \mathbb{P}_\mu\left(|U_N(a) - \mu^{-1}(a)| > \sqrt{x}, \mathcal{E}_N\right) dx \\ &= \int_0^\infty 2y \mathbb{P}_\mu\left(|U_N(a) - \mu^{-1}(a)| > y, \mathcal{E}_N\right) dy \\ &\leq \int_0^\infty 2y \left(\frac{C}{Ny^3} \wedge 1\right) dy. \end{aligned}$$

For the last inequality, we used (6.1) together with the fact that a probability cannot be larger than one. Hence,

$$\begin{aligned} \mathbb{E}_\mu\left(|U_N(a) - \mu^{-1}(a)|^2 \mathbb{1}_{\mathcal{E}_N}\right) &\leq \int_0^{N^{-1/3}} 2y dy + \int_{N^{-1/3}}^\infty \frac{2C}{Ny^2} dy \\ &\leq N^{-2/3}\left(1 + 2C\right). \end{aligned}$$

Combining with (6.5) yields

$$\mathbb{E}_\mu\left(|U_N(a)-\mu^{-1}(a)|^2\right) \quad \leq \quad N^{-2/3}\,(2+2C)\,,$$

which completes the proof of the Theorem (by taking C to be $2+2C$) where C is the constant from Lemma (6.2). □

7. Proof of Theorem 4.1

In the sequel we denote $\theta_{0K} = [K\theta_0]/K$.

Lemma 7.1 (Lemma 6.1 equivalent)**.** *Under Condition (S), there exist $c>0$ and $N_0>0$ such that for all $N \geq N_0$, we have*

$$F_X(u+r_{0K}) - F_X(\theta_{0K}+r_{0K}) - F_X(u-r_{0K}) + F_X(\theta_{0K}-r_{0K}) \leq -c(u-\theta_0)^2 \tag{7.1}$$

for all $f_X \in \mathcal{F}_1$ with corresponding primitive F_X, $u \in [0,1]$ and $|u-\theta_0| \geq N^{-1/3}$.

Proof of Lemma 7.1. By definition, both $\theta_N - r_{0K}$ and $\theta_N + r_{0K}$ belong to the set $\{\bar{x}_0,\ldots,\bar{x}_K\}$ and we have

$$|r_{0K}-r_0| \leq K^{-1}. \tag{7.2}$$

Moreover, all $f_X \in \mathcal{F}_1$ are bounded in supremum norm by C_2 so we have

$$\begin{aligned} &F_X(u+r_{0K}) - F_X\left(\theta_{0K}+r_{0K}\right) - F_X(u-r_{0K}) + F_X\left(\theta_{0K}-r_{0K}\right) \\ &\quad = F_X(u+r_0) - F_X\left(\theta_0+r_0\right) - F_X(u-r_0) + F_X\left(\theta_0-r_0\right) + o(1) \end{aligned}$$

uniformly for all u such that $|u-\theta_0| \geq a/2$. Hence, it follows from (4.5) that we can find $c>0$ such that

$$\begin{aligned} \sup_{|u-\theta_0|\geq a/2} &\left\{F_X(u+r_{0K}) - F_X\left(\theta_{0K}+r_{0K}\right) - F_X(u-r_{0K}) + F_X\left(\theta_{0K}-r_{0K}\right)\right\} \\ &< -ca^2 \end{aligned}$$

for sufficiently large N, which proves that the inequality in (7.1) holds for all u with $|u-\theta_0| \geq a/2$.

Now, consider u such that $N^{-1/3} \leq |u-\theta_0| \leq a/2$. It follows from the Taylor expansion that

$$\begin{aligned} &F_X(u+r_{0K}) - F_X\left(\theta_{0K}+r_{0K}\right) - F_X(u-r_{0K}) + F_X\left(\theta_{0K}-r_{0K}\right) \\ &\quad = (u-\theta_{0K})\,f_X\left(\theta_{0K}+r_{0K}\right) + \frac{(u-\theta_{0K})^2}{2} f_X'(\xi_1) \\ &\qquad -(u-\theta_{0K}) f_X\left(\theta_{0K}-r_{0K}\right) - \frac{(u-\theta_{0K})^2}{2} f_X'(\xi_2) \end{aligned}$$

where ξ_1 and ξ_2 depend on u and are such that $|\xi_1-(\theta_0+r_0)| \le a/2+2K^{-1} \le a$ and similarly, $|\xi_2-(\theta_0-r_0)| \le a$. Using (4.2) and (4.3), we arrive at

$$\begin{aligned} &F_X(u+r_{0K}) - F_X(\theta_{0K}+r_{0K}) - F_X(u-r_{0K}) + F_X(\theta_{0K}-r_{0K}) \\ &\quad \le (u-\theta_{0K}) f_X(\theta_{0K}+r_{0K}) - (u-\theta_{0K}) f_X(\theta_{0K}-r_{0K}) - \frac{(u-\theta_{0K})^2}{6}\epsilon \end{aligned}$$

for all u with $N^{-1/3} \le |u-\theta_0| \le a/2$. Since $N^{-1/3} \gg K^{-1}$ we conclude that for sufficiently large N and all u with $N^{-1/3} \le |u-\theta_0| \le a/2$,

$$\begin{aligned} &F_X(u+r_{0K}) - F_X(\theta_{0K}+r_{0K}) - F_X(u-r_{0K}) + F_X(\theta_{0K}-r_{0K}) \\ &\quad \le (u-\theta_{0K}) f_X(\theta_0+r_0) - (u-\theta_{0K}) f_X(\theta_0-r_0) - \frac{(u-\theta_0)^2}{12}\epsilon. \end{aligned}$$

By (4.5), θ_0 maximizes $\theta \mapsto F_X(\theta+r_0) - F_X(\theta-r_0)$ and the maximum is achieved in the open interval $(r_0, 1-r_0)$ so the derivative vanishes at θ_0:

$$f_X(\theta_0+r_0) - f_X(\theta_0-r_0) = 0. \tag{7.3}$$

Combining the previous two displays yields the result. □

Lemma 7.2. *There exists $C>0$ and $N_0>0$ such that for all $f_X \in \mathcal{F}_1$ with corresponding primitive F_X, all $x > N^{-1/3}$, $u_0 \in [0,1]$, and $N \ge N_0$,*

$$\mathbb{E}_{f_X}\left(\sup_{|u-u_0|\le x} |F_N(u) - F_N(u_0) - F_X(u) + F_X(u_0)|^2\right) \le \frac{Cx}{N}. \tag{7.4}$$

The proof is available in the Appendix.

In the sequel, we denote by $\mathbb{P}_f$ the underlying probability when the distributions of the observations are such that the true density f_X of F_X defined in (3.1) is equal to f.

Lemma 7.3. *There exists $C>0$ and $N_0>0$ such that for all $f \in \mathcal{F}_1$, $a \in \mathbb{R}$, $x>0$, and $N \ge N_0$*

$$\mathbb{P}_f(|\theta_N - \theta_0| \ge x) \le \frac{C}{Nx^3}. \tag{7.5}$$

Proof of Lemma 7.3. The inequality in the lemma is obvious for $x \in (0, N^{-1/3})$ since for such x's, it suffices to choose $C \ge 1$ so that the right hand side is larger than one. Hence, in the sequel we consider $x \ge N^{-1/3}$. It follows from the definition of θ_N and Lemma 7.1 that the event $\{\theta_N - \theta_0 \ge x\}$ is included in the event that there exists $u \in [0,1]$ such that both $u-r_{0K}$ and $u+r_{0K}$ belong to the set $\{\bar{x}_0, \ldots, \bar{x}_K\}$, $u-\theta_0 \ge x$ and

$$\begin{aligned} &\Lambda_N(u+r_{0K}) - \Lambda_N(u-r_{0K}) - \Lambda_N(\theta_{0K}+r_{0K}) + \Lambda_N(\theta_{0K}-r_{0K}) \\ &\quad -F_X(u+r_{0K}) + F_X(u-r_{0K}) + F_X(\theta_{0K}+r_{0K}) - F_X(\theta_{0K}-r_{0K}) \\ &\qquad \ge c(u-\theta_0)^2. \end{aligned}$$

As Λ_N matches F_N on the set $\{\bar{x}_0, \ldots, \bar{x}_K\}$, the function Λ_N can be replaced by F_N in the above display. We have (7.2) where $K^{-1} \ll N^{-1/3} \leq x$ and therefore, we can assume without loss of generality that $|r_{0K} - r_0| \leq x/2$. Hence, the event $\{\theta_N - \theta_0 \geq x\}$ is included in the event that there exist $u, v \in [0,1]$ such that $u - \theta_0 \geq x/2$, $v - \theta_0 \geq x/2$ and

$$\begin{aligned} &F_N(u + r_0) - F_N(v - r_0) - F_N(\theta_{0K} + r_{0K}) + F_N(\theta_{0K} - r_{0K}) \\ &\quad - F_X(u + r_0) + F_X(v - r_0) + F_X(\theta_{0K} + r_{0K}) - F_X(\theta_{0K} - r_{0K}) \\ &\quad \geq c(u - \theta_0)^2. \end{aligned}$$

This implies that the event $\{\theta_N - \theta_0 \geq x\}$ is included in the event that there exist $u, v \in [0,1]$ such that $u - \theta_0 \geq x/2$, $v - \theta_0 \geq x/2$ and either

$$\begin{aligned} &F_N(u + r_0) - F_N(\theta_{0K} + r_{0K}) - F_X(u + r_0) + F_X(\theta_{0K} + r_{0K}) \\ &\quad \geq \frac{c}{2}(u - \theta_0)^2 \end{aligned}$$

or

$$\begin{aligned} &-F_N(v - r_0) + F_N(\theta_{0K} - r_{0K}) + F_X(v - r_0) - F_X(\theta_{0K} - r_{0K}) \\ &\quad \geq \frac{c}{2}(u - \theta_0)^2. \end{aligned}$$

Hence,

$$\mathbb{P}_f(\theta_N - \theta_0 \geq x) \leq \sum_{k \geq 0} \mathbb{P}_f(\mathcal{A}_k) + \sum_{k \geq 0} \mathbb{P}_f(\mathcal{B}_k) \tag{7.6}$$

where for all $k \geq 0$, $\mathcal{A}_k$ is the event that there exist $u \in [0,1]$ such that $u - \theta_0 \in [x2^k/2, x2^{k+1}/2]$, and

$$\begin{aligned} &F_N(u + r_0) - F_N(\theta_{0K} + r_{0K}) - F_X(u + r_0) + F_X(\theta_{0K} + r_{0K}) \\ &\quad \geq \frac{c}{2}x^2 2^{2(k-1)} \end{aligned}$$

and $\mathcal{B}_k$ is the event that there exist $v \in [0,1]$ such that $v - \theta_0 \in [x2^k/2, x2^{k+1}/2]$, and

$$\begin{aligned} &-F_N(v - r_0) + F_N(\theta_{0K} - r_{0K}) + F_X(v - r_0) - F_X(\theta_{0K} - r_{0K}) \\ &\quad \geq \frac{c}{2}x^2 2^{2(k-1)}. \end{aligned}$$

We will deal with the first sum in the right hand side of (7.6), the second sum being similar. Since $|\theta_{0K} - r_{0K} - (\theta_0 - r_0)| \leq 2K^{-1}$ where $K^{-1} \ll N^{-1/3} \leq x$, we have for all $k \geq 0$ that $\mathbb{P}_f(\mathcal{A}_k)$ is bounded from above by

$$\begin{aligned} 2\mathbb{P}_f\Bigg(\sup_{u \leq \theta_0 + x2^{k+1}} \Big\{ &F_N(u + r_0) - F_N(\theta_0 + r_0) - F_X(u + r_0) + F_X(\theta_0 + r_0) \Big\} \\ &\geq \frac{c}{4}x^2 2^{2(k-1)} \Bigg). \end{aligned}$$

Combining this with Lemma 7.2 and the Markov inequality, we conclude that

$$\mathbb{P}_f\left(\mathcal{A}_k\right) \leq \frac{32Cx2^{k+1}N}{c^2x^4 2^{4(k-1)}} = \frac{C2^{10-3k}}{Nc^2x^3}.$$

Since $\sum_{k\geq 0} 2^{-3k}$ is finite, we conclude that there exists $C > 0$ such that

$$\sum_{k\geq 0} \mathbb{P}_f\left(\mathcal{A}_k\right) \leq \frac{C}{Nx^3}.$$

Similar arguments show that the same inequality holds with $\mathcal{A}_k$ replaced by $\mathcal{B}_k$ so we conclude from (7.6) that there exists $C > 0$ such that

$$\mathbb{P}_f\left(\theta_N - \theta_0 \geq x\right) \leq \frac{C}{Nx^3}.$$

It can be proved similarly that

$$\mathbb{P}_f\left(\theta_N - \theta_0 \leq -x\right) \leq \frac{C}{Nx^3},$$

and the lemma follows. □

Proof of Theorem 4.1. It follows from the Fubini theorem that for all $f \in \mathcal{F}_1$ and $N \geq N_0$,

$$\begin{aligned}
\mathbb{E}_f\left(|\theta_N - \theta_0|^2\right) &= \int_0^\infty \mathbb{P}_f\left(|\theta_N - \theta_0| > \sqrt{x}\right) dx \\
&= \int_0^\infty 2y\mathbb{P}_f\left(|\theta_N - \theta_0| > y\right) dy \\
&\leq \int_0^\infty 2y\left(\frac{C}{Ny^3} \wedge 1\right) dy.
\end{aligned}$$

For the last inequality, we used (7.5) together with the fact that a probability cannot be larger than one. Hence,

$$\begin{aligned}
\mathbb{E}_f\left(|\theta_N - \theta_0|^2\right) &\leq \int_0^{N^{-1/3}} 2ydy + \int_{N^{-1/3}}^\infty \frac{2C}{Ny^2} dy \\
&\leq N^{-2/3}\left(1 + 2C\right),
\end{aligned}$$

which completes the proof of the Theorem (by taking C to be $1 + 2C$). □

Acknowledgement

We thank Professor Ya'acov Ritov for some fruitful discussions (with the first author) that inspired the construction of the estimates proposed in this paper. We are also grateful to Debarghya Mukherjee for help with simulations.

Appendix

A.1. Preparatory lemmas

Lemma A.1. *Assume that the distribution function F_X taken from (3.1) has a density function f_X on $[0,1]$ that satisfies (3.2) for some positive numbers C_1, C_2. Let F_N be the empirical distribution function taken from (2.2) and let F_N^{-1} be the corresponding empirical quantile function. We then have*

$$\mathbb{P}\left(\sup_{t\in[0,1]}|F_N(t)-F_X(t)|>x\right) \leq 2\sum_{j=1}^{m}\exp(-2n_jx^2) \tag{A.1}$$

and

$$\mathbb{P}\left(\sup_{t\in[0,1]}|F_N^{-1}(t)-F_X^{-1}(t)|>x\right) \leq 4\sum_{j=1}^{m}\exp(-2n_jC_1^2x^2) \tag{A.2}$$

for all N and $x>0$.

Proof. Let F_{X_j} denote the common distribution function of the X_i's from sample j and denote by (X_{ji}, Y_{ji}), $i=1,\ldots,n_j$ the observations from sample j. It follows from the triangle inequality that

$$\sup_{t\in[0,1]}|F_N(t)-F_X(t)| \leq \sum_{j=1}^{m}\frac{n_j}{N}\sup_{t\in[0,1]}\left|\frac{1}{n_j}\sum_{i=1}^{n_j}\mathbb{1}_{X_{ji}\leq t}-F_{X_j}(t)\right|$$

where we recall that $\sum_{j=1}^m n_j = N$. Hence, for all $x>0$ we have

$$\begin{aligned}
&\mathbb{P}\left(\sup_{t\in[0,1]}|F_N(t)-F_X(t)|>x\right)\\
&\quad\leq \mathbb{P}\left(\sup_{t\in[0,1]}\left|\frac{1}{n_j}\sum_{i=1}^{n_j}\mathbb{1}_{X_{ji}\leq t}-F_{X_j}(t)\right|>x \text{ for some } j\in\{1,\ldots,m\}\right)\\
&\quad\leq \sum_{j=1}^{m}\mathbb{P}\left(\sup_{t\in[0,1]}\left|\frac{1}{n_j}\sum_{i=1}^{n_j}\mathbb{1}_{X_{ji}\leq t}-F_{X_j}(t)\right|>x\right).
\end{aligned}$$

Since for all fixed j, the random variables X_{ji}, $i=1,\ldots,n_j$ are i.i.d. with distribution function F_{X_j}, it follows from Corollary 1 in [15] that

$$\mathbb{P}\left(\sup_{t\in[0,1]}\left|\frac{1}{n_j}\sum_{i=1}^{n_j}\mathbb{1}_{X_{ji}\leq t}-F_{X_j}(t)\right|>x\right)\leq 2\exp(-2n_jx^2).$$

Combining the two preceding displays completes the proof of (A.1).

Now, consider (A.2). Since f_X is supported on $[0,1]$, both F_N^{-1} and F_X^{-1} take values in $[0,1]$ so the sup-distance between those functions is less than or equal to one. This means that the probability on the left hand side of (A.2) is equal to zero for all $x\geq 1$. Hence, it suffices to prove (A.2) for $x\in(0,1)$. As is customary, we use the notation $y_+=\max(y,0)$ and $y_-=-\min(y,0)$ for all real numbers$y$. This means that $|y|=\max(y_-,y_+)$. Recall the switching relation

for the empirical distribution and empirical quantile functions: for arbitrary $a \in [0,1]$ and $t \in [0,1]$, we have

$$F_N(a) \geq t \Longleftrightarrow a \geq F_N^{-1}(t). \tag{A.3}$$

For all $x \in (0,1)$ we then have

$$\begin{aligned}\mathbb{P}\left(\sup_{t\in[0,1]} (F_N^{-1}(t) - F_X^{-1}(t))_+ > x\right) &= \mathbb{P}\left(\exists t \in [0,1] :\ F_N^{-1}(t) > x + F_X^{-1}(t)\right) \\ &= \mathbb{P}\left(\exists t \in [0,1] :\ t > F_N(x + F_X^{-1}(t))\right).\end{aligned}$$

Using $t = F_X(F_X^{-1}(t))$ together with the change of variable $u = x + F_X^{-1}(t)$ we obtain

$$\begin{aligned}\mathbb{P}\left(\sup_{t\in[0,1]} (F_N^{-1}(t) - F_X^{-1}(t))_+ > x\right) &\leq \mathbb{P}\left(\exists u \geq x :\ F_X(u-x) > F_N(u)\right) \\ &= \mathbb{P}\left(\exists u \in (x,1) :\ F_X(u-x) > F_N(u)\right).\end{aligned}$$

For the last equality, we use the fact that $F_X(u-x) \leq 1 = F_N(u)$ for all $u \geq 1$, and $F_X(u-x) = 0 \leq F_N(u)$ for all $u \leq x$. With C_1 taken from (3.2) we have $F_X(u-x) < F_X(u) - C_1 x$ for all $x \in (0,1)$ and $u \in (x,1)$. Combining this to the previous display yields

$$\begin{aligned}\mathbb{P}\left(\sup_{t\in[0,1]} (F_N^{-1}(t) - F_X^{-1}(t))_+ > x\right) &\leq \mathbb{P}\left(\exists u \in (x,1) :\ F_X(u) - F_N(u) > C_1 x\right) \\ &\leq \mathbb{P}\left(\sup_{u\in\mathbb{R}} |F_X(u) - F_N(u)| > C_1 x\right) \\ &\leq 2\sum_{j=1}^{m} \exp(-2n_j C_1^2 x^2).\end{aligned} \tag{A.4}$$

For the last inequality, we used (A.1). On the other hand, for all $x \in (0,1)$ we have

$$\begin{aligned}\mathbb{P}\left(\sup_{t\in[0,1]} (F_N^{-1}(t) - F_X^{-1}(t))_- > x\right) &\leq \mathbb{P}\left(\exists t \in [0,1] :\ F_N^{-1}(t) < F_X^{-1}(t) - x\right) \\ &\leq \mathbb{P}\left(\exists u \in (x,1) :\ F_N^{-1}(F_X(u)) \leq u - x\right),\end{aligned}$$

using the change of variable $u = F_X^{-1}(t)$. Hence, with the switching relation we obtain

$$\begin{aligned}&\mathbb{P}\left(\sup_{t\in[0,1]} (F_N^{-1}(t) - F_X^{-1}(t))_- > x\right) \\ &\quad\leq \mathbb{P}\left(\exists u \in (x,1) :\ F_X(u) \leq F_N(u-x)\right) \\ &\quad\leq \mathbb{P}\left(\exists u \in (x,1) :\ F_X(u-x) + C_1 x < F_N(u-x)\right),\end{aligned}$$

using that $F_X(u-x) < F_X(u) - C_1 x$ for all $x \in (0,1)$ and $u \in (x,1)$. Using again (A.1) together with the change of variable $v = u - x$, we arrive at

$$\begin{aligned}\mathbb{P}\left(\sup_{t\in[0,1]}(F_N^{-1}(t) - F_X^{-1}(t))_- > x\right) &\leq \mathbb{P}\left(\sup_{v\in\mathbb{R}}|F_X(v) - F_N(v)| > C_1 x\right) \\ &\leq 2\sum_{j=1}^{m}\exp(-2n_j C_1^2 x^2).\end{aligned}$$

Combining the previous display with (A.4) completes the proof of (A.2) since $|y| \leq y_- + y_+$ for all $y \in \mathbb{R}$. □

Lemma A.2. *Under the assumptions of Theorem 3.3, for all $p > 0$ there exists $K_p > 0$ such that for all N,*

$$\mathbb{E}\left(\sup_{t\in[0,1]}|F_N(t) - F_X(t)|^p\right) \leq K_p N^{-p/2}. \tag{A.5}$$

Proof. It follows from the Fubini theorem that

$$\begin{aligned}\mathbb{E}\left(\sup_{t\in[0,1]}|F_N(t) - F_X(t)|^p\right) &= \int_0^\infty \mathbb{P}\left(\sup_{t\in[0,1]}|F_N(t) - F_X(t)|^p > x\right)dx \\ &= \int_0^\infty px^{p-1}\mathbb{P}\left(\sup_{t\in[0,1]}|F_N(t) - F_X(t)| > x\right)dx.\end{aligned}$$

Combining this with (A.1) and the fact that a probability cannot be larger than one then yields

$$\begin{aligned}&\mathbb{E}\left(\sup_{t\in[0,1]}|F_N(t) - F_X(t)|^p\right) \\ &\quad\leq N^{-p/3} + 2\sum_{j=1}^{m}\int_{N^{-1/3}}^\infty px^{p-1}\exp(-2n_j x^2)dx \\ &\quad\leq N^{-p/3} + 2N\int_{N^{-1/3}}^\infty px^{p-1}\exp(-2N^{2/3}(\log N)^3 x^2)dx\end{aligned}$$

for sufficiently large N, where we used (3.4) for the last inequality. The result follows by computing the integral on the right-hand side. □

A.2. Proof of Theorem 3.2

Denote by $\widetilde{F}_N$ the step function on $[0,1]$ such that $\widetilde{F}_N(\overline{x}_k) = F_N(\overline{x}_k)$ for all $k = 0,\ldots,K$, and $\widetilde{F}_N$ is constant on all intervals $[\overline{x}_{k-1}, \overline{x}_k)$ for $k = 1,\ldots,K$. We denote by $\widetilde{F}_N^{-1}$ the corresponding inverse function:

$$\widetilde{F}_N^{-1}(t) = \inf\{x \in [0,1] \text{ such that } \widetilde{F}_N(x) \geq t\}.$$

Since $\widetilde{F}_N^{-1}\circ\widetilde{F}_N(\overline{x}_k)=\overline{x}_k$ for all $k=0,\dots,K$, it follows from the characterization in (2.3) that

$$U_N(a)=\widetilde{F}_N^{-1}(V_N(a)) \tag{A.6}$$

for all $a\in\mathbb{R}$, where

$$V_N(a)=\underset{u\in\{\widetilde{F}_N(\overline{x}_0),\dots,\widetilde{F}_N(\overline{x}_K)\}}{\operatorname{argmax}}\{\Lambda_N\circ\widetilde{F}_N^{-1}(u)-au\}.$$

The following lemma provides tail bound probabilities for V_N.

Lemma A.3. *With $\varepsilon_i=Y_i-\mu(X_i)$ for all $i=1,\dots,N$, assume that there exists $\sigma>0$ such that $\mathbb{E}[\varepsilon_i^p|X_i]\le\sigma^p$ for all i and some $p\ge 2$, with probablity one. Assume that F_X has a density function f_X on $[0,1]$ that satisfies (3.2) for some positive numbers C_1,C_2. Then, there exists $C>0$ that depends only on p,C_2 and σ such that*

$$\mathbb{P}\left(V_N(a)\ge x\right)\le\frac{C}{N^{p/2}x^{p-1}(a-\mu(0))^p}$$

for all $a>\mu(0)$ and

$$\mathbb{P}\left(1-V_N(a)\ge x\right)\le\frac{C}{N^{p/2}x^{p-1}(\mu(1)-a)^p}.$$

for all $a<\mu(1)$.

A proof of this lemma follows the proof the main theorem.

Proof of Theorem 3.2. Similar to the proof of Theorem 3.1 for the inverse problem, we would like to restrict ourselves to the event $\mathcal{E}_N$ from Lemma 6.1, where θ can be chosen arbitrarily large. However, we do not have an analogue of (6.5) for the direct problem since $\widehat{\mu}_N$ is not bounded as is U_N. Hence, we first prove that $\widehat{\mu}_N$ remains bounded by a power of N apart possibly on a negligible set. For this task, consider an arbitrary $A>0$ such that $A+\mu(0)>0$, and note that for all $t\in[0,1]$, and all non-increasing functions μ on $[0,1]$, we have

$$\mathbb{E}_\mu\left[\widehat{\mu}_N^2(t)\mathbb{1}_{\widehat{\mu}_N(t)>A+\mu(0)}\right]\le\mathbb{E}_\mu\left[\widehat{\mu}_N^2(0)\mathbb{1}_{\widehat{\mu}_N(0)>A+\mu(0)}\right].$$

Hence, it follows from the Fubini theorem that for all non-increasing $\mu\in\mathcal{F}_1$,

$$\begin{aligned}&\mathbb{E}_\mu\left[\widehat{\mu}_N^2(t)\mathbb{1}_{\widehat{\mu}_N(t)>A+\mu(0)}\right]\\&\quad\le\int_0^\infty\mathbb{P}_\mu(\widehat{\mu}_n(0)\mathbb{1}_{\widehat{\mu}_N(0)>A+\mu(0)}>\sqrt{y})dy\\&\quad=(A+\mu(0))^2\mathbb{P}_\mu(\widehat{\mu}_N(0)>A+\mu(0))+\int_{A+\mu(0)}^\infty 2y\mathbb{P}_\mu(\widehat{\mu}_N(0)>y)dy.\end{aligned}$$

Note that if $\widehat{\mu}_N(0)>y$ for some $y\in\mathbb{R}$, then for the inverse we must have $U_N(y)>0$. Since U_N can only assume values in the set of jump points of $\widehat{\mu}_N$ it is of the form $\overline{x}_k=k/K$ for some $k\ge 1$. Next, $\widehat{\mu}_N$ can have jumps only at those $\overline{x}_k$ where $\widetilde{F}_N$ has a jump, i.e. $\widetilde{F}_N(\overline{x}_k)>\widetilde{F}_N(\overline{x}_{k-1})$. Since the size of a jump of $\widetilde{F}_N$ is at least N^{-1}, we must have $\widetilde{F}_N(\overline{x}_k)\ge N^{-1}$ and therefore,

$F_N(\overline{x}_k) = \widetilde{F}_N(\overline{x}_k) \geq N^{-1}$. Thus,

$$V_N(y) = \widetilde{F}_N(U_N(y)) = F_N(U_N(y)) = F_N(\overline{x}_k) \geq N^{-1},$$

implying that for all $\mu \in \mathcal{F}_1$,

$$\begin{aligned}
&\mathbb{E}_\mu\left[\widehat{\mu}_N^2(t)\mathbb{1}_{\widehat{\mu}_N(t)>A+\mu(0)}\right] \\
&\quad\leq (A+\mu(0))^2\mathbb{P}_\mu(V_N(A+\mu(0)) \geq N^{-1}) + \int_{A+\mu(0)}^{\infty} 2y\mathbb{P}_\mu(V_N(y) \geq N^{-1})dy.
\end{aligned}$$

With C taken from Lemma (A.3) where it is assumed that $p > 2$, we arrive at

$$\begin{aligned}
&\mathbb{E}_\mu\left[\widehat{\mu}_N^2(t)\mathbb{1}_{\widehat{\mu}_N(t)>A+\mu(0)}\right] \\
&\quad\leq CN^{-1+p/2}(A+\mu(0))^2A^{-p} + 2CN^{-1+p/2}\int_{A+\mu(0)}^{\infty} y(y-\mu(0))^{-p}dy \\
&\quad= CN^{-1+p/2}(A+\mu(0))^2A^{-p} + 2CN^{-1+p/2}\left\{\frac{A^{2-p}}{p-2} + \mu(0)\frac{A^{1-p}}{p-1}\right\} \\
&\quad\leq CN^{-1+p/2}(A+C_5)^2A^{-p} + 2CN^{-1+p/2}\left\{\frac{A^{2-p}}{p-2} + C_5\frac{A^{1-p}}{p-1}\right\}.
\end{aligned}$$

With $A = N^{(3p-2)/(6(p-2))}$, this proves that there exists $C' > 0$ that depends only on σ, p, C_2 and C_5 such that

$$\mathbb{E}_\mu\left[(\widehat{\mu}_N(t))^2\mathbb{1}_{\widehat{\mu}_N(t)>A+\mu(0)}\right] \quad\leq\quad C'N^{-2/3}$$

for all $t \in [0,1]$ and $\mu \in \mathcal{F}_1$. Now, with $A = N^{(3p-2)/(6(p-2))}$,

$$\begin{aligned}
&\mathbb{E}_\mu\left[(\widehat{\mu}_N(t)-\mu(t))^2\mathbb{1}_{\widehat{\mu}_N(t)>A+\mu(0)}\right] \\
&\quad\leq \mathbb{E}_\mu\left[2\left(\widehat{\mu}_N^2(t)+\mu^2(t)\right)\mathbb{1}_{\widehat{\mu}_N(t)>A+\mu(0)}\right] \\
&\quad\leq 2C'N^{-2/3} + 2\max\{|\mu(0)|,|\mu(1)|\}^2\mathbb{P}\left(\widehat{\mu}_N(0) > A+\mu(0)\right) \\
&\quad\leq 2C'N^{-2/3} + 2\max\{|\mu(0)|,|\mu(1)|\}^2\mathbb{P}\left(V_N(A+\mu(0)) \geq N^{-1}\right),
\end{aligned}$$

similar as above, whence

$$\begin{aligned}
\mathbb{E}_\mu\left[(\widehat{\mu}_N(t)-\mu(t))^2\mathbb{1}_{\widehat{\mu}_N(t)>A+\mu(0)}\right] &\leq 2C'N^{-2/3} + 2C{C_5}^2N^{-1+p/2}A^{-p} \\
&\leq C''N^{-2/3}
\end{aligned}$$

where C'' depends only on σ, p, C_2, and C_5. This enables us to restrict to the event $\mathcal{E}_N$ of Lemma 6.1, provided that θ is chosen sufficiently large in the lemma. Indeed, with $\theta > (5p-6)/(3(p-2))$, the previous inequality implies that with $A = N^{(3p-2)/(6(p-2))}$ and N sufficiently large,

$$\begin{aligned}
\mathbb{E}_\mu\left[\{(\widehat{\mu}_N(t)-\mu(t))_+\}^2\mathbb{1}_{\overline{\mathcal{E}}_N}\right] &\leq (A+\mu(0)-\mu(t))^2\mathbb{P}_\mu(\overline{\mathcal{E}}_N) + C''N^{-2/3} \\
&\leq (A+2C_5)^2\mathbb{P}_\mu(\overline{\mathcal{E}}_N) + C''N^{-2/3} \\
&\leq 2C''N^{-2/3}
\end{aligned}$$

for all $t \in [0,1]$ and all $\mu \in \mathcal{F}_1$. It can be shown similarly that for N sufficiently large,

$$\mathbb{E}_\mu \left[\{(\widehat{\mu}_N(t) - \mu(t))_-\}^2 \mathbb{1}_{\overline{\mathcal{E}}_N} \right] \quad \leq \quad 2C'' N^{-2/3}$$

for all $t \in [0,1]$ and all $\mu \in \mathcal{F}_1$, implying that

$$\limsup_{N\to\infty} \sup_{\mu\in\mathcal{F}_1} N^{2/3} \mathbb{E}_\mu \left[(\widehat{\mu}_N(t) - \mu(t))^2 \mathbb{1}_{\overline{\mathcal{E}}_N} \right] \quad \leq \quad 4C''$$

for all $t \in [0,1]$. Hence, it now suffices to prove that there exists $C > 0$ that depends only on $\sigma, p, C_1, C_2, C_3, C_4, C_5, \delta$ such that

$$\limsup_{N\to\infty} \sup_{\mu\in\mathcal{F}_1} N^{2/3} \mathbb{E}_\mu \left[(\widehat{\mu}_N(t) - \mu(t))^2 \mathbb{1}_{\mathcal{E}_N} \right] \leq C. \tag{A.7}$$

To prove this, fix $\mu \in \mathcal{F}_1$ arbitrarily, and invoke the Fubini Theorem to obtain that

$$\begin{aligned} \mathbb{E}_\mu \left[\{(\widehat{\mu}_N(t) - \mu(t))_+\}^2 \mathbb{1}_{\mathcal{E}_N} \right] &= \int_0^\infty 2y \mathbb{P}_\mu \left(\widehat{\mu}_N(t) - \mu(t) \geq y, \mathcal{E}_N \right) dy \\ &= \int_0^\infty 2y \mathbb{P}_\mu \left(U_N(\mu(t) + y) \geq t, \mathcal{E}_N \right) dy, \end{aligned} \tag{A.8}$$

using the switch relation (2.4) for the last equality. We split the above integral into the sum of two integrals and first consider

$$I_1 = \int_0^{\mu(0)-\mu(t)} 2y \mathbb{P}_\mu \left(U_N(\mu(t) + y) \geq t, \mathcal{E}_N \right) dy.$$

With C_4 taken from the definition of $\mathcal{F}_1$ we have

$$t = \mu^{-1}(\mu(t)) \geq \mu^{-1}(\mu(t) + y) + yC_4^{-1}$$

for all $t \in [0,1]$ and $y \in [0, \mu(0) - \mu(t)]$. Combining Lemma 6.2 with the fact that a probability cannot be larger than one then yields

$$\begin{aligned} I_1 &\leq N^{-2/3} + \int_{N^{-1/3}}^{\mu(0)-\mu(t)} 2y \mathbb{P}_\mu \left(U_N(\mu(t) + y) - \mu^{-1}(\mu(t) + y) \geq yC_4^{-1}, \mathcal{E}_N \right) dy \\ &\leq N^{-2/3} + \int_{N^{-1/3}}^{\infty} \frac{2CC_4^3}{Ny^2} dy \\ &\leq N^{-2/3} \left(1 + 2CC_4^3 \right). \end{aligned} \tag{A.9}$$

Next, Lemma 6.2 yields

$$\begin{aligned} I_2 &:= \int_{\mu(0)-\mu(t)}^{\infty} 2y \mathbb{P}_\mu \left(U_N(\mu(t) + y) \geq t, \mathcal{E}_N \right) dy \\ &\leq \int_{\mu(0)-\mu(t)}^{\mu(0)-\mu(t)+N^{1/6}} 2y \mathbb{P}_\mu \left(U_N(0) \geq t, \mathcal{E}_N \right) dy \end{aligned}$$

$$
\begin{aligned}
&\quad + \int_{\mu(0)-\mu(t)+N^{1/6}}^{\infty} 2y\mathbb{P}_\mu\left(U_N(\mu(t)+y) \geq t, \mathcal{E}_N\right) dy \\
&\leq \frac{C}{Nt^3} \int_{\mu(0)-\mu(t)}^{\mu(0)-\mu(t)+N^{1/6}} 2y dy \\
&\quad + \int_{\mu(0)-\mu(t)+N^{1/6}}^{\infty} 2y\mathbb{P}_\mu\left(U_N(\mu(t)+y) \geq t, \mathcal{E}_N\right) dy,
\end{aligned}
$$

where the first term on the right-hand side is equal to

$$
\begin{aligned}
&\frac{C}{Nt^3}\left((\mu(0)-\mu(t)+N^{1/6})^2 - (\mu(0)-\mu(t))^2\right) \\
&\quad = \frac{C}{Nt^3}\left(2(\mu(0)-\mu(t))N^{1/6} + N^{1/3}\right) \\
&\quad \leq \frac{C}{Nt^3}(4C_5 N^{-1/6}+1)N^{1/3} \\
&\quad \leq \frac{2C}{\delta^3} N^{-2/3}
\end{aligned}
$$

for sufficiently large N, for all $t \geq \delta$ and $\mu \in \mathcal{F}_1$. Using the connection (A.6) between U_N and V_N yields

$$
I_2 \leq \frac{2C}{\delta^3} N^{-2/3} + \int_{\mu(0)-\mu(t)+N^{1/6}}^{\infty} 2y\mathbb{P}_\mu\left(V_N(\mu(t)+y) \geq \widetilde{F}_N(t), \mathcal{E}_N\right) dy,
$$

where $\widetilde{F}_N(t) = \widetilde{F}_N([Kt]K^{-1}) = F_N([Kt]K^{-1})$ by definition of $\widetilde{F}_N$ and F_N. Regarding the proof of Lemma 6.1, it can be seen that on $\mathcal{E}_N$ we have

$$
\sup_{t\in[0,1]} |F_N(t) - F_X(t)| \leq C_2 N^{-1/3}
$$

whence

$$
\begin{aligned}
I_2 &\leq \frac{2C}{\delta^3} N^{-2/3} \\
&\quad + \int_{\mu(0)-\mu(t)+N^{1/6}}^{\infty} 2y\mathbb{P}_\mu\left(V_N(\mu(t)+y) \geq F_X([Kt]K^{-1}) - C_2 N^{-1/3}\right) dy \\
&\leq \frac{2C}{\delta^3} N^{-2/3} \\
&\quad + \int_{\mu(0)-\mu(t)+N^{1/6}}^{\infty} 2y\mathbb{P}_\mu\left(V_N(\mu(t)+y) \geq C_1(t-K^{-1}) - C_2 N^{-1/3}\right) dy \\
&\leq \frac{2C}{\delta^3} N^{-2/3} + \int_{\mu(0)-\mu(t)+N^{1/6}}^{\infty} 2y\mathbb{P}_\mu\left(V_N(\mu(t)+y) \geq C_1\delta/2\right) dy,
\end{aligned}
$$

for all $t \in [\delta, 1-\delta]$, provided that N is sufficiently large. Hence, it follows from Lemma A.3 where it is assumed that $p > 2$, that

$$
I_2 \quad \leq \quad \frac{2C}{\delta^3} N^{-2/3} + \frac{2^p C}{(C_1\delta)^{p-1}} \int_{\mu(0)-\mu(t)+N^{1/6}}^{\infty} \frac{yN^{-p/2}}{(y+\mu(t)-\mu(0))^p} dy.
$$

For the integral on the right-hand side we have

$$\begin{aligned}
&\int_{\mu(0)-\mu(t)+N^{1/6}}^{\infty} \frac{yN^{-p/2}}{(y+\mu(t)-\mu(0))^p}dy \\
&\quad= \int_{\mu(0)-\mu(t)+N^{1/6}}^{\infty} \frac{N^{-p/2}}{(y+\mu(t)-\mu(0))^{p-1}}dy \\
&\qquad+ \int_{\mu(0)-\mu(t)+N^{1/6}}^{\infty} \frac{(\mu(0)-\mu(t))N^{-p/2}}{(y+\mu(t)-\mu(0))^p}dy \\
&\quad= \int_{N^{1/6}}^{\infty} \frac{N^{-p/2}}{u^{p-1}}du + \int_{N^{1/6}}^{\infty} \frac{(\mu(0)-\mu(t))N^{-p/2}}{u^p}du \\
&\quad\le \frac{1}{p-2}N^{(1-2p)/3} + \frac{2C_5}{p-1}N^{(1-4p)/6}.
\end{aligned}$$

Hence, we can find $\tilde{C}$ that depends only on $p, \sigma, C_1 - C_5$ such that

$$I_2 \le \tilde{C}N^{-2/3}$$

for sufficiently large N, for all $\mu \in \mathcal{F}_1$ and $t \in [\delta, 1-\delta]$.

Combining this with (A.9) and (A.8) proves that there exists $C > 0$ that depends only on $\sigma, p, C_1, C_2, C_3, C_4, C_5, \delta$ such that

$$\limsup_{N\to\infty} \sup_{\mu\in\mathcal{F}_1} N^{2/3}\mathbb{E}_\mu\left[(\widehat{\mu}_N(t)-\mu(t))_+^2 \mathbb{1}_{\mathcal{E}_N}\right] \le C.$$

It can be proved similarly that

$$\limsup_{N\to\infty} \sup_{\mu\in\mathcal{F}_1} N^{2/3}\mathbb{E}_\mu\left[(\widehat{\mu}_N(t)-\mu(t))_-^2 \mathbb{1}_{\mathcal{E}_N}\right] \le C,$$

which completes the proof (A.7), and hence the proof of the theorem. □

Proof of Lemma A.3. For all $a \notin [\mu(1), \mu(0)]$ and $u \in \{\overline{x}_0, \ldots, \overline{x}_K\}$, define $e(a,u)$ as in (A.21). We then have (A.22) where

$$\frac{[Kg(a)]}{K} = g(a) = \begin{cases} 0 & \text{if } a > \mu(0) \\ 1 & \text{if } a < \mu(1). \end{cases}$$

and f is given by (A.23). Note that (A.24) is no longer true for $a \notin [\mu(1), \mu(0)]$ since in such a case, $a \ne \mu \circ g(a)$. Instead, we will use

$$\begin{aligned}
&\left(\mu\left(\frac{[Kg(a)]}{K}\right) - a\right)\left(F_N(u) - F_N\left(\frac{[Kg(a)]}{K}\right)\right) \\
&\qquad= \begin{cases} (\mu(0)-a)F_N(u) & \text{if } a > \mu(0) \\ (\mu(1)-a)(F_N(u)-1) & \text{if } a < \mu(1), \end{cases}
\end{aligned}$$

using that $F_N(0) = 0$ and $F_N(1) = 1$. Since $f(a,u) \ge 0$ for all a, u (A.22) yields

$$e(a,u) \le \begin{cases} (\mu(0)-a)F_N(u) & \text{if } a > \mu(0) \\ (\mu(1)-a)(F_N(u)-1) & \text{if } a < \mu(1) \end{cases} \tag{A.10}$$

for all $a \notin [\mu(1), \mu(0)]$ and $u \in \{\overline{x}_0, \ldots, \overline{x}_K\}$.

Since $\Lambda_N(\overline{x}_0) - a\widetilde{F}_N(\overline{x}_0) = 0$, it follows from the definition of V_N that the following inequalities hold for all $x > 0$ and $a > \mu(0)$:

$$\begin{aligned}
&\mathbb{P}\left(V_N(a) \geq x\right) \\
&\quad \leq \mathbb{P}\left(\max_{u \in \{\widetilde{F}_N(\overline{x}_0), \ldots, \widetilde{F}_N(\overline{x}_K)\},\ u \geq x} \{\Lambda_N \circ \widetilde{F}_N^{-1}(u) - au\} \geq 0 \right) \\
&\quad = \mathbb{P}\left(\max_{u \in \{\widetilde{F}_N(\overline{x}_0), \ldots, \widetilde{F}_N(\overline{x}_K)\},\ u \geq x} \{M_N \circ \widetilde{F}_N^{-1}(u) + e(a, \widetilde{F}_N^{-1}(u))\} \geq 0 \right)
\end{aligned}$$

where $M_N(u) = \Lambda_N(u) - \mathbb{E}^X(\Lambda_N(u))$ takes the form (6.3). The first inequality in (A.10) then yields

$$\begin{aligned}
&\mathbb{P}\left(V_N(a) \geq x\right) \\
&\quad \leq \mathbb{P}\left(\max_{u \in \{\widetilde{F}_N(\overline{x}_0), \ldots, \widetilde{F}_N(\overline{x}_K)\},\ u \geq x} \{M_N \circ \widetilde{F}_N^{-1}(u) + (\mu(0) - a)u\} \geq 0 \right) \\
&\quad \leq \sum_{k \geq 0} \mathbb{P}\left(\max_{u \in \{\widetilde{F}_N(\overline{x}_0), \ldots, \widetilde{F}_N(\overline{x}_K)\},\ u \in [x2^k, x2^{k+1}]} \{M_N \circ \widetilde{F}_N^{-1}(u)\} \geq (a - \mu(0))x2^k \right).
\end{aligned}$$

Let $p \geq 2$ and $\sigma > 0$ such that $\mathbb{E}[\varepsilon_i^p | X_i] \leq \sigma^p$ for all i, almost surely. The process M_n is a centered martingale under $\mathbb{P}^X$ which, according to Theorem 3 in [18], satisfies

$$\begin{aligned}
\mathbb{E}^X\left|M_N(u)\right|^p &\leq \frac{A_p}{N^p} \max\left\{ \sum_{i=1}^N \mathbb{E}^X|\varepsilon_i|^p \mathbb{1}_{X_i \leq u}; \left(\sum_{i=1}^N \mathbb{E}^X|\varepsilon_i|^2 \mathbb{1}_{X_i \leq u} \right)^{p/2} \right\} \\
&\leq \frac{A_p \sigma^p}{N^p} \max\left\{ NF_N(u); (NF_N(u))^{p/2} \right\} \\
&\leq \frac{A_p \sigma^p F_N(u)}{N^{p/2}}
\end{aligned}$$

for all $u \in [0,1]$ and $A_p = (p/2)^{p/2} 2^{p + p^2/4}$. For the penultimate inequality, we used that $\mathbb{E}^X|\varepsilon_i|^2 \leq (\mathbb{E}^X|\varepsilon_i|^p)^{2/p}$ thanks to the Holder inequality whereas for the last inequality, we used that $N \leq N^{p/2}$ and $F_N^{p/2}(u) \leq F_N(u)$. Combining the two preceding displays with the Doob inequality yields that for all $x > 0$, $\mathbb{P}\left(V_N(a) \geq x\right)$ is less than or equal to

$$\begin{aligned}
&\sum_{k \geq 0} \mathbb{E}\left[\mathbb{P}^X\left(\max_{u \in \{\widetilde{F}_N(\overline{x}_0), \ldots, \widetilde{F}_N(\overline{x}_K)\},\ u \in [x2^k, x2^{k+1}]} \{M_N \circ \widetilde{F}_N^{-1}(u)\} \geq (a - \mu(0))x2^k \right) \right] \\
&\quad \leq \sum_{k \geq 0} \mathbb{E}\left[\frac{A_p \sigma^p F_N(x2^{k+1})}{N^{p/2}(a - \mu(0))^p (x2^k)^p} \right].
\end{aligned}$$

With C_2 taken from (3.2) we conclude that for all $x > 0$,

$$\begin{aligned}\mathbb{P}\left(V_N(a) \geq x\right) &\leq \sum_{k\geq 0} \frac{A_p \sigma^p F_X(x2^{k+1})}{N^{p/2}(a-\mu(0))^p (x2^k)^p} \\ &\leq \sum_{k\geq 0} \frac{2A_p C_2 \sigma^p}{N^{p/2}(a-\mu(0))^p (x2^k)^{p-1}}.\end{aligned}$$

Since $C := 2A_p C_2 \sigma^p \sum_{k\geq 0} 2^{-k(p-1)}$ is finite, we conclude that

$$\mathbb{P}\left(V_N(a) \geq x\right) \leq \frac{C}{N^{p/2}(a-\mu(0))^p x^{p-1}},$$

which proves the first assertion. For the second assertion, since $\overline{x}_K = \widetilde{F}_N(\overline{x}_K) = 1$, we write for $a < \mu(1)$ and $x > 0$ that $\mathbb{P}\left(1 - V_N(a) \geq x\right)$ is less than or equal to

$$\begin{aligned}&\mathbb{P}\Bigg(\max_{u\in\{\widetilde{F}_N(\overline{x}_0),\ldots,\widetilde{F}_N(\overline{x}_K)\},\ 1-u\geq x} \{\Lambda_N \circ \widetilde{F}_N^{-1}(u) - au\} \geq \Lambda_N(1) - a\Bigg) \\ &= \mathbb{P}\Bigg(\max_{u\in\{\widetilde{F}_N(\overline{x}_0),\ldots,\widetilde{F}_N(\overline{x}_K)\},\ 1-u\geq x} \{M_N \circ \widetilde{F}_N^{-1}(u) - M_N(1) + e(a, \widetilde{F}_N^{-1}(u))\} \geq 0\Bigg).\end{aligned}$$

The first inequality in (A.10) then yields that $\mathbb{P}\left(1 - V_N(a) \geq x\right)$ is less than or equal to

$$\begin{aligned}&\mathbb{P}\Bigg(\max_{u\in\{\widetilde{F}_N(\overline{x}_0),\ldots,\widetilde{F}_N(\overline{x}_K)\}, 1-u\geq x} \{M_N \circ \widetilde{F}_N^{-1}(u) - M_N(1) - (\mu(1)-a)(1-u)\} \geq 0\Bigg) \\ &\quad \leq \sum_{k\geq 0} \mathbb{P}\Bigg(\max_{u\in\{\widetilde{F}_N(\overline{x}_0),\ldots,\widetilde{F}_N(\overline{x}_K)\}, 1-u\leq x2^{k+1}} \{M_N \circ \widetilde{F}_N^{-1}(u) - M_N(1)\} \\ &\qquad\qquad \geq (\mu(1)-a)x2^k\Bigg),\end{aligned}$$

and we use the Doob inequality, similar as above. Details are omitted. □

A.3. Proof of Theorem 3.3

It follows from (2.3) together with Lemma 6.2 that with probablity tending to one,

$$N^{1/3}(U_N(a) - g(a)) = \operatorname*{argmax}_{u\in H_N} \{\Lambda_N(g(a) + N^{-1/3}u) - aF_N(g(a) + N^{-1/3}u)\}$$

where H_N is the set of all $u \in \mathbb{R}$ such that $g(a) + N^{-1/3}u \in \{\overline{x}_0, \ldots, \overline{x}_K\}$ and $|u| \leq v_N$, where v_N is an arbitrary sequence that diverges to infinity as $N \to \infty$. In the sequel, we consider a sequence v_N such that $v_N \leq \log N$ for all N. Hence, with probablity that tends to one, $N^{1/3}(U_N(a) - g(a))$ is the location of the

maximum over $u \in H_N$ of

$$N^{2/3}\left(M_N(g(a)+N^{-1/3}u) - M_N\left(\frac{[Kg(a)]}{K}\right)\right) + N^{2/3}e(a, g(a)+N^{-1/3}u)$$

where $M_N(u) = \Lambda_N(u) - E^X(\Lambda_N(u))$ for all $u \in \{\overline{x}_0, \ldots, \overline{x}_K\}$ and e is taken from (A.21), that is

$$e(a, g(a)+N^{-1/3}u) = \frac{1}{N}\sum_{i=1}^{N}\mu(X_i)\left(\mathbb{1}_{X_i \le g(a)+N^{-1/3}u} - \mathbb{1}_{X_i \le [Kg(a)]K^{-1}}\right) - a\left(F_N(g(a)+N^{-1/3}u) - F_N\left(\frac{[Kg(a)]}{K}\right)\right). \quad \text{(A.11)}$$

We extend M_N and $e(a, \ .\ )$ as constant functions in between two consecutive points in H_N so that

$$\begin{aligned} &N^{1/3}(U_N(a) - g(a)) \\ &\quad = \operatorname*{argmax}_{|u| \le v_N}\left\{N^{2/3}\left(M_N(g(a)+N^{-1/3}u) - M_N\left(\frac{[Kg(a)]}{K}\right)\right)\right. \\ &\qquad \left. + N^{2/3}e(a, g(a)+N^{-1/3}u)\right\} + o_p(1). \end{aligned} \quad \text{(A.12)}$$

Now, since $a = \mu(t) + N^{-1/3}x$ for some fixed $x \in \mathbb{R}$ and $t \in (0,1)$, and $g' = 1/\mu' \circ g$ on $(\mu(1), \mu(0))$ is bounded by assumption, we have

$$g(a) = t + O(N^{-1/3}). \quad \text{(A.13)}$$

Hence, for sufficiently large N, every X_i that lies between $[Kg(a)]K^{-1}$ and $g(a)+N^{-1/3}u$ for some $|u| \le v_N$ also lies in $[t - N^{-1/3}\log N, t + N^{-1/3}\log N]$. This implies that for all such X_i's there exists $\theta_i \in [t - N^{-1/3}\log N, t + N^{-1/3}\log N]$ such that

$$\begin{aligned} \mu(X_i) &= \mu\left(\frac{[Kg(a)]}{K}\right) + \left(X_i - \frac{[Kg(a)]}{K}\right)\mu'(\theta_i) \\ &= \mu\left(\frac{[Kg(a)]}{K}\right) + \left(X_i - \frac{[Kg(a)]}{K}\right)(\mu'(t) + o(1)) \end{aligned} \quad \text{(A.14)}$$

where the small o-term is uniform, by continuity of μ' over the compact interval $[t - N^{-1/3}\log N, t + N^{-1/3}\log N]$. Plugging this in (A.11), and using the notation f in (A.23), yields

$$\begin{aligned} &e(a, g(a)+N^{-1/3}u) \\ &\quad = (\mu'(t) + o(1))\, f(a, g(a)+N^{-1/3}u) \\ &\qquad + \left(\mu\left(\frac{[Kg(a)]}{K}\right) - a\right)\left(F_N(g(a)+N^{-1/3}u) - F_N\left(\frac{[Kg(a)]}{K}\right)\right). \end{aligned}$$

It can be seen from the proof of Lemma 6.1 that (A.26) holds on the event $\mathcal{E}_N$, whose probability tends to one as $N \to \infty$, implying that

$$\begin{aligned} & F_N(g(a) + N^{-1/3}u) - F_N\left(\frac{[Kg(a)]}{K}\right) \\ &= F_X(g(a) + N^{-1/3}u) - F_X\left(\frac{[Kg(a)]}{K}\right) + O_p(N^{-1/3}(\log N)^{-1}) \\ &= O_p(N^{-1/3}v_N + K^{-1} + N^{-1/3}(\log N)^{-1}) \\ &= O_p(N^{-1/3}v_N) \end{aligned}$$

uniformy over $u \in H_N$. Since (A.24) holds for all $a \in [\mu(1), \mu(0)]$, combining the two preceding displays yields

$$e(a, g(a) + N^{-1/3}u) = (\mu'(t) + o(1))\, f(a, g(a) + N^{-1/3}u) + O_p(K^{-1}N^{-1/3}v_N).$$

Next, we invoke (A.31), that holds on the event $\mathcal{E}_N$ uniformly over a and u, to conclude that

$$\begin{aligned} & e(a, g(a) + N^{-1/3}u) \\ &\quad = (\mu'(t) + o(1)) \int_{g(a)}^{g(a)+N^{-1/3}u} (z - g(a))\, f_X(z)dz + o_p(N^{-2/3}) \end{aligned}$$

uniformly over $u \in H_N$, provided that $v_N \ll \min\{\log N; N^{-1/3}K\}$. By assumption, $N^{-1/3}K$ diverges to infinity as $N \to \infty$, so we can find a sequence v_N that satisfies the above condition and that diverges to infinity as $N \to \infty$, as required in the definition of H_N. In the sequel, we consider a sequence v_N that satisfies the above conditions and in addition, the below condition:

$$v_N \ll \left(\max\left\{\sup_{|z-t|\le N^{-1/3}\log N} |f_X(z) - f_\infty(z)|, \sup_{|z-t|\le N^{-1/3}\log N} |f_\infty(t) - f_\infty(z)|\right\}\right)^{-1/2}. \tag{A.15}$$

Note that by assumption, the right-hand side of the inequality in the above display diverges to infinity as $N \to \infty$, which ensures existence of such a sequence v_N. We then have

$$\begin{aligned} & e(a, g(a) + N^{-1/3}u) \\ &\quad = (\mu'(t) + o(1)) \int_{g(a)}^{g(a)+N^{-1/3}u} (z - g(a))\, f_\infty(z)dz + o_p(N^{-2/3}), \end{aligned}$$

using that for $u \geq 0$ (and similarly for $u \leq 0$),

$$\begin{aligned}
&\left| \int_{g(a)}^{g(a)+N^{-1/3}u} (z-g(a))\,(f_X(z)-f_\infty(z))dz \right| \\
&\quad \leq \int_{g(a)}^{g(a)+N^{-1/3}u} (z-g(a))\,|f_X(z)-f_\infty(z)|dz \\
&\quad \leq \frac{N^{-2/3}v_N^2}{2} \sup_{|z-g(a)|\leq N^{-1/3}v_N} |f_X(z)-f_\infty(z)|
\end{aligned}$$

uniformly for all $|u| \leq v_N$, which implies

$$\begin{aligned}
&\left| \int_{g(a)}^{g(a)+N^{-1/3}u} (z-g(a))\,(f_X(z)-f_\infty(z))dz \right| \\
&\quad \leq \frac{N^{-2/3}v_N^2}{2} \sup_{|z-t|\leq N^{-1/3}\log N} |f_X(z)-f_\infty(z)| \\
&\quad = o(N^{-2/3})
\end{aligned}$$

thanks to (A.13), (A.15) and the assumption that $v_n \ll \log N$. Similarly,

$$\begin{aligned}
&\left| \int_{g(a)}^{g(a)+N^{-1/3}u} (z-g(a))\,(f_\infty(z)-f_\infty(t))dz \right| \\
&\quad \leq \frac{N^{-2/3}v_N^2}{2} \sup_{|z-g(a)|\leq N^{-1/3}v_N} |f_\infty(z)-f_\infty(t)| \\
&\quad \leq \frac{N^{-2/3}v_N^2}{2} \sup_{|z-t|\leq N^{-1/3}\log N} |f_\infty(z)-f_\infty(t)| \\
&\quad = o(N^{-2/3})
\end{aligned}$$

and therefore,

$$\begin{aligned}
&e(a, g(a)+N^{-1/3}u) \\
&\quad = (\mu'(t)+o(1)) \int_{g(a)}^{g(a)+N^{-1/3}u} (z-g(a))\,f_\infty(z)dz + o_p(N^{-2/3}) \\
&\quad = (\mu'(t)+o(1))\,f_\infty(t) \int_{g(a)}^{g(a)+N^{-1/3}u} (z-g(a))\,dz + o_p(N^{-2/3}).
\end{aligned}$$

Hence we obtain

$$N^{2/3}e(a, g(a)+N^{-1/3}u) = -(|\mu'(t)|+o(1))f_\infty(t)\frac{u^2}{2} + o_p(1). \tag{A.16}$$

On the other hand, with

$$Z_N(u) = N^{2/3}\left(M_N(g(a)+N^{-1/3}u) - M_N\left(\frac{[Kg(a)]}{K}\right)\right);$$

where M_N is as defined in (6.3) for all $u \in \{\overline{x}_0, \ldots, \overline{x}_K\}$, we have

$$Z_N(u) = N^{-1/3} \sum_{j=1}^{m} \sum_{i=1}^{n_j} \varepsilon_{ji} \left(\mathbb{1}_{X_{ji} \leq g(a) + N^{-1/3} u} - \mathbb{1}_{X_{ji} \leq [Kg(a)]K^{-1}} \right) \tag{A.17}$$

where we denote by (X_{ji}, Y_{ji}), $i = 1, \ldots, n_j$ the observations from sample j, for $j = 1, \ldots, m$, and $\varepsilon_{ji} = Y_{ji} - \mu(X_{ji})$. Note that the process Z_N is centered and has been extended to $\mathbb{R}$ by being constant in between two consecutive points in H_N. For all $u \geq v \geq 0$ in H_N we have

$$\begin{aligned} &N^{2/3} \mathbb{E}\left[Z_N(u) Z_N(v) \right] \\ &\quad = \sum_{j=1}^{m} \sum_{i=1}^{n_j} \mathbb{E}\left[\varepsilon_{ji}^2 \mathbb{1}_{[Kg(a)]K^{-1} < X_{ji} \leq g(a) + N^{-1/3} u} \mathbb{1}_{[Kg(a)]K^{-1} < X_{ji} \leq g(a) + N^{-1/3} v} \right] \\ &\quad = \sum_{j=1}^{m} \sum_{i=1}^{n_j} \mathbb{E}\left[\sigma_j^2(X_{ji}) \mathbb{1}_{[Kg(a)]K^{-1} < X_{ji} \leq g(a) + N^{-1/3} v} \right], \end{aligned}$$

where the last equality is obtained by conditioning with respect to X_{ji} and using that $u \geq v \geq 0$. With u, v fixed, this implies that

$$\mathbb{E}\left[Z_N(u) Z_N(v) \right] = N^{-2/3} \sum_{j=1}^{m} \sum_{i=1}^{n_j} \mathbb{E}\left[\sigma_j^2(t) \mathbb{1}_{[Kg(a)]K^{-1} < X_{ji} \leq g(a) + N^{-1/3} v} \right] + o(1)$$

using that for $u, v \in H_N$

$$\begin{aligned} &\left| \mathbb{E}\left[Z_N(u) Z_N(v) \right] - N^{-2/3} \sum_{j=1}^{m} \sum_{i=1}^{n_j} \mathbb{E}\left[\sigma_j^2(t) \mathbb{1}_{[Kg(a)]K^{-1} < X_{ji} \leq g(a) + N^{-1/3} v} \right] \right| \\ &\qquad \leq N^{-2/3} \sum_{j=1}^{m} \sum_{i=1}^{n_j} \mathbb{E}\left[|\sigma_j^2(X_{ji}) - \sigma_j^2(t)| \mathbb{1}_{[Kg(a)]K^{-1} < X_{ji} \leq g(a) + N^{-1/3} v} \right] \\ &\qquad \leq N^{-2/3} \omega(N^{-1/3} \log N) \sum_{j=1}^{m} \sum_{i=1}^{n_j} \mathbb{E}\left[\mathbb{1}_{[Kg(a)]K^{-1} < X_{ji} \leq g(a) + N^{-1/3} v} \right] \end{aligned}$$

where $\omega(\delta) \to 0$ as $\delta \to 0$ by assumption, and

$$\begin{aligned} &N^{-2/3} \sum_{j=1}^{m} \sum_{i=1}^{n_j} \mathbb{E}\left[\mathbb{1}_{[Kg(a)]K^{-1} < X_{ji} \leq g(a) + N^{-1/3} v} \right] \\ &\qquad = N^{1/3} \left| F_X(g(a) + N^{-1/3} v) - F_X([Kg(a)]K^{-1}) \right| \\ &\qquad = O(1). \end{aligned}$$

Hence,

$$\begin{aligned}
&\mathbb{E}\left[Z_N(u)Z_N(v)\right] \\
&\quad = N^{-2/3}\sum_{j=1}^{m}\sum_{i=1}^{n_j}\sigma_j^2(t)\mathbb{P}\left([Kg(a)]K^{-1} < X_{ji} \le g(a) + N^{-1/3}v\right) + o(1) \\
&\quad = N^{-2/3}\sum_{j=1}^{m} n_j\sigma_j^2(t)\int_{[Kg(a)]K^{-1}}^{g(a)+N^{-1/3}v} f_j(z)dz + o(1)
\end{aligned}$$

for all fixed real numbers $u \ge v \ge 0$. It follows that

$$\begin{aligned}
&\left|\mathbb{E}\left[Z_N(u)Z_N(v)\right] - N^{-2/3}\sum_{j=1}^{m} n_j\sigma_j^2(t)\int_{[Kg(a)]K^{-1}}^{g(a)+N^{-1/3}v} f_j(t)dz\right| \\
&\quad \le N^{-2/3}\sum_{j=1}^{m} n_j\sigma_j^2(t)\omega(N^{-1/3}\log N)\left(N^{-1/3}v + O(K^{-1})\right) + o(1) \\
&\quad \le o(1)N^{-1}\sum_{j=1}^{m} n_j\sigma_j^2(t) + o(1),
\end{aligned}$$

since $\omega(\delta) \to 0$ as $\delta \to 0$. The Jensen inequality for conditional expectation combined with Assumption $\tilde{A}_4$ shows that $\sigma_j^2(t) \le \sigma^2$ for all i and t and therefore, $N^{-1}\sum_{j=1}^{m} n_j\sigma_j^2(t) \le \sigma^2$. This implies that

$$\begin{aligned}
\mathbb{E}\left[Z_N(u)Z_N(v)\right] &= N^{-2/3}\sum_{j=1}^{m} n_j\sigma_j^2(t)f_j(t)(N^{-1/3}v + o(N^{-1/3})) + o(1) \\
&= \sigma_X^2(t)v + o(1).
\end{aligned}$$

We conclude that for all $u \ge v \ge 0$, $\mathbb{E}\left[Z_N(u)Z_N(v)\right] = cov(Z_N(u), Z_N(v))$ converges to $\sigma_\infty^2(t)v$. The case of negative u and v can be treated likewise and therefore, $cov(Z_N(u), Z_N(v))$ converges to $\sigma_\infty^2(t)(|u| \wedge |v|)$ if $uv \ge 0$. It can be seen similarly that it converges to zero if $uv < 0$ (hence u and v have different signs). Hence, the covariance converges to $\sigma_\infty(t)cov(W(u), W(v))$, so we conclude from the Lindeberg-Feller theorem that jointly,

$$(Z_N(u_1), \ldots, Z_N(u_k)) \to_d \sigma_\infty(t)(W(u_1), \ldots, W(u_k)) \tag{A.18}$$

for all fixed $u_1, \ldots, u_k \in \mathbb{R}$, as $N \to \infty$. Now, consider the restriction of Z_N to the compact interval $[-M, M]$, for a fixed $M > 0$. For all $\delta > 0$ and $\epsilon > 0$ we have

$$\begin{aligned}
&\mathbb{P}\left(\sup_{|t-s|\le\delta;\ s,t\in[-M,M]} |Z_N(s) - Z_N(t)| \ge \epsilon\right) \\
&\quad \le \sum_{k=-M[\delta^{-1}]-1}^{M[\delta^{-1}]} \mathbb{P}\left(2\sup_{|t-k\delta|\le 2\delta} |Z_N(k\delta) - Z_N(t)| \ge \epsilon\right).
\end{aligned} \tag{A.19}$$

Let π be the permutation such that the $X_{\pi(j)}$ are ordered in j, that is $X_{\pi(1)} < \cdots < X_{\pi(N)}$ a.s. Let $\mathbb{P}_X$ denote the conditional probability given $X_1, \ldots, X_N$. Since $\varepsilon_{\pi(1)}, \ldots, \varepsilon_{\pi(N)}$ are centered and independent under $\mathbb{P}_X$, the process $\{Z_N(k\delta) - Z_N(t),\ t \geq k\delta\}$ is a forward centered martingale whereas $\{Z_N(k\delta) - Z_N(t),\ t \leq k\delta\}$ is a reverse centered martingale conditionally on $X_1, \ldots, X_N$, for all k. Hence, it follows from the Doob inequality that for all k,

$$\mathbb{P}\left(2 \sup_{|t-k\delta|\leq 2\delta} |Z_N(k\delta) - Z_N(t)| \geq \epsilon\right) \leq \frac{2^p}{\epsilon^p} \left(\mathbb{E}|Z_N(k\delta) - Z_N((k-2)\delta)|^p + \mathbb{E}|Z_N(k\delta) - Z_N((k+2)\delta)|^p\right). \tag{A.20}$$

Note that the inequalities above are first obtained for the conditional probabilities and then integrated over the distribution of X for the unconditional. Now, it follows from the Rosenthal inequality, see [18], that for all k and a constant C that depends only on p, we have

$$\mathbb{E}|Z_N(k\delta) - Z_N((k+2)\delta)|^p \leq CN^{-p/3} \left(\sum_{i=1}^N \mathbb{E}(|\varepsilon_i|^p \mathbb{1}_{X_i \in I_k}) + \left(\sum_{i=1}^N \mathbb{E}(|\varepsilon_i|^2 \mathbb{1}_{X_i \in I_k})\right)^{p/2}\right).$$

Here, $I_k = (g(a)+N^{-1/3}k\delta, g(a)+N^{-1/3}(k+2)\delta]$ (at least if $g(a)+N^{-1/3}k\delta, g(a)$ and $N^{-1/3}(k+2)\delta$ both belong to H_N) and p is taken from Assumption $\tilde{A}_4$. Hence, with f_X taken from (3.5) we have

$$\begin{aligned}
&\mathbb{E}|Z_N(k\delta) - Z_N((k+2)\delta)|^p \\
&\quad \leq C\sigma^p N^{-p/3} \left(\sum_{i=1}^N \mathbb{E}(\mathbb{1}_{X_i \in I_k}) + \left(\sum_{i=1}^N \mathbb{E}(\mathbb{1}_{X_i \in I_k})\right)^{p/2}\right) \\
&\quad = C\sigma^p N^{-p/3} \left(N \int_{I_k} f_X(u)du + N^{p/2} \left[\int_{I_k} f_X(u)du\right]^{p/2}\right).
\end{aligned}$$

It follows from the Assumption $\tilde{A}_1$ that f_X is bounded by a constant A that does not depend on N and therefore,

$$\begin{aligned}
&\mathbb{E}|Z_N(k\delta) - Z_N((k+2)\delta)|^p \\
&\quad \leq C\sigma^p N^{-p/3} \left(2AN^{2/3}\delta + \left[2AN^{2/3}\delta\right]^{p/2}\right)(1 + o(1)) \\
&\quad \leq 2C\sigma^p N^{-p/3} \left[2AN^{2/3}\delta\right]^{p/2} \\
&\quad = 2C\sigma^p \left[2A\delta\right]^{p/2}
\end{aligned}$$

for N sufficiently large. Arguing similarly for $\mathbb{E}|Z_N(k\delta) - Z_N((k-2)\delta)|^p$ we conclude from (A.20) that there exists $C > 0$ that depends only on p and A

such that

$$\mathbb{P}\left(2\sup_{|t-k\delta|\le 2\delta}|Z_N(k\delta)-Z_N(t)|\ge\epsilon\right) \le C\sigma^p\epsilon^{-p}\delta^{p/2}$$

for all k. Summing up this inequality over all k on the right-hand side of (A.19), we obtain that there exists $C>0$ that depends only on p and A such that for all $\delta>0$ and $\epsilon>0$,

$$\mathbb{P}\left(\sup_{|t-s|\le\delta\ ;\ s,t\in[-M,M]}|Z_N(s)-Z_N(t)|\ge\epsilon\right)\le CM\sigma^p\epsilon^{-p}\delta^{-1+p/2}.$$

Since $p>2$, this converges to zero as $\delta\to 0$. Using (A.18), it follows from [7, Theorem 7.5] that Z_N converges weakly to $\sigma_\infty W$ on all compact intervals $[-M,M]$. Combining this with (A.12) and (A.16) we conclude that $N^{1/3}(U_N(a)-g(a))$ is the location of the maximum of a process that weakly converges to the continuous Gaussian process

$$\sigma_\infty(t)W(u)-\frac{|\mu'(t)|f_\infty(t)}{2}u^2,\ u\in\mathbb{R}.$$

The above process achieves its maximum at a unique point $\mathbb{T}$ by Lemma 2.6 of [11], and it follows from Lemma 6.2 that $N^{1/3}(U_N(a)-g(a))$ is uniformly tight. Hence, Corollary 5.58 in van der Vaart shows that $N^{1/3}(U_N(a)-g(a))$ converges in distribution to $\mathbb{T}$. Now, $\mathbb{T}$ is also the unique location of the maximum of the process

$$W(u)-\frac{|\mu'(t)|f_\infty(t)}{2\sigma_\infty(t)}u^2,\ u\in\mathbb{R}.$$

Changing scale in the Brownian motion finally shows that

$$\left(\frac{|\mu'(t)|f_\infty(t)}{2\sigma_\infty(t)}\right)^{2/3}\mathbb{T}$$

has the same distribution as $\mathbb{Z}$, which completes the proof. □

A.4. Proof of Lemma 6.1

For all $a\in\mathbb{R}$ and $u\in\{\overline{x}_0,\dots,\overline{x}_K\}$ such that $|u-g(a)|\ge N^{-1/3}$, define

$$e(a,u)=E^X\Lambda_N(u)-E^X\Lambda_N\left(\frac{[Kg(a)]}{K}\right)-a\left(F_N(u)-F_N\left(\frac{[Kg(a)]}{K}\right)\right). \tag{A.21}$$

By definition of Λ_N we have

$$e(a,u)=\frac{1}{N}\sum_{i=1}^N\mu(X_i)\left(\mathbb{1}_{X_i\le u}-\mathbb{1}_{X_i\le[Kg(a)]K^{-1}}\right)-a\left(F_N(u)-F_N\left(\frac{[Kg(a)]}{K}\right)\right).$$

Now, $X_i \neq [Kg(a)]K^{-1}$ for all i, almost surely since X_i has a continuous distribution function, so (3.3) implies that

$$\left|\mu(X_i) - \mu\left(\frac{[Kg(a)]}{K}\right)\right| \geq \left|X_i - \frac{[Kg(a)]}{K}\right| C_3,$$

implying that

$$e(a,u) \leq \left(\mu\left(\frac{[Kg(a)]}{K}\right) - a\right)\left(F_N(u) - F_N\left(\frac{[Kg(a)]}{K}\right)\right) - C_3 f(a,u) \quad \text{(A.22)}$$

with a decreasing function μ, where

$$f(a,u) = \frac{1}{N}\sum_{i=1}^{N}\left(X_i - \frac{[Kg(a)]}{K}\right)\left(\mathbb{1}_{X_i \leq u} - \mathbb{1}_{X_i \leq [Kg(a)]K^{-1}}\right). \quad \text{(A.23)}$$

Using again (3.3), we obtain that for all $a \in [\mu(1), \mu(0)]$,

$$\begin{aligned}\left|\mu\left(\frac{[Kg(a)]}{K}\right) - a\right| &= \left|\mu\left(\frac{[Kg(a)]}{K}\right) - \mu \circ g(a)\right| \\ &\leq C_4 K^{-1}. \end{aligned} \quad \text{(A.24)}$$

On the other hand, since F_X has a bounded derivative that satisfies (3.2) we have

$$\begin{aligned}&\left|F_X(u) - F_X\left(\frac{[Kg(a)]}{K}\right)\right| \\ &\quad \leq |F_X(u) - F_X(g(a))| + \left|F_X(g(a)) - F_X\left(\frac{[Kg(a)]}{K}\right)\right| \\ &\quad \leq C_2\left(|u - g(a)| + K^{-1}\right) \\ &\quad \leq 2C_2|u - g(a)| \end{aligned} \quad \text{(A.25)}$$

for sufficiently large N, using that $K^{-1} = o(N^{-1/3})$ whereas $|u - g(a)| \geq N^{-1/3}$ for the last inequality. Next, since $m \leq N$, it follows from (A.1) in the Appendix that

$$\mathbb{P}\left(\sup_{t\in[0,1]} |F_N(t) - F_X(t)| > x\right) \leq 2N \exp\left(-2x^2 \min_{1\leq j\leq m} n_j\right)$$

for all $x > 0$. With (3.4), we obtain

$$\mathbb{P}\left(\sup_{t\in[0,1]} |F_N(t) - F_X(t)| > x\right) \leq 2N \exp\left(-2x^2 N\lambda\right)$$

for all $x > 0$. With $\widetilde{\mathcal{E}}_N$ the event that

$$\sup_{t\in[0,1]} |F_N(t) - F_X(t)| \leq C_2 N^{-1/3}(\log N)^{-1} \quad \text{(A.26)}$$

we conclude from the previous display that

$$\begin{aligned} 1-\mathbb{P}(\widetilde{\mathcal{E}}_N) &\leq 2N\exp\left(-2C_2^2N^{1/3}\lambda(\log N)^{-2}\right) \\ &\ll N^{-\theta}, \end{aligned} \tag{A.27}$$

where we used (3.4) for the last claim. Combining (A.22), (A.24) and (A.25) proves that on $\widetilde{\mathcal{E}}_N$, we have

$$\begin{aligned} &e(a,u) \\ &\leq C_4K^{-1}\left(\left|F_X(u)-F_X\left(\frac{[Kg(a)]}{K}\right)\right|+2\sup_{t\in[0,1]}|F_N(t)-F_X(t)|\right)-C_3f(a,u) \\ &\leq 2C_2C_4K^{-1}(|u-g(a)|+N^{-1/3})-C_3f(a,u) \end{aligned} \tag{A.28}$$

for all $a\in[\mu(1),\mu(0)]$. The inequality in (A.22) holds also for $a\notin[\mu(1),\mu(0)]$, and in that case,

$$\frac{[Kg(a)]}{K}=g(a)=\begin{cases}0 & \text{if } a>\mu(0)\\ 1 & \text{if } a<\mu(1),\end{cases}$$

implying that

$$\left(\mu\left(\frac{[Kg(a)]}{K}\right)-a\right)\left(F_N(u)-F_N\left(\frac{[Kg(a)]}{K}\right)\right)\leq 0.$$

Hence, the inequality in (A.28) holds for all $a\in\mathbb{R}$ and $u\in\{\overline{x}_0,\ldots,\overline{x}_K\}$. Using that $K^{-1}=o(N^{-1/3})$ whereas $|u-g(a)|\geq N^{-1/3}$, we conclude that on $\widetilde{\mathcal{E}}_N$,

$$e(a,u) \leq o(u-g(a))^2-C_3f(a,u)$$

uniformly over all a and u such that $|u-g(a)|\geq N^{-1/3}$. Hence, it suffices to prove that with $f(a,u)$ taken from (A.23), there exists $\tilde{c}>0$ that only depends on C_1 such that on an event $\mathcal{E}_N$ whose probability is larger than $1-N^{-\theta}$, and such that $\mathcal{E}_N\subset\widetilde{\mathcal{E}}_N$, we have

$$\begin{aligned} f(a,u)\geq\tilde{c}(u-g(a))^2 &\text{ for all } a\in\mathbb{R},\ u\in\{\overline{x}_0,\ldots,\overline{x}_K\} \\ &\text{such that } |u-g(a)|\geq N^{-1/3}. \end{aligned} \tag{A.29}$$

Similar to (A.27), if follows from (A.2) in the Appendix that

$$\begin{aligned} \mathbb{P}\left(\sup_{x\in[0,1]}|F_N^{-1}(x)-F_X^{-1}(x)|>\frac{N^{-1/3}}{\log N}\right) &\leq 4N\exp\left(-2C_1^2N^{1/3}(\log N)^{-2}\lambda\right) \\ &\ll N^{-\theta}. \end{aligned} \tag{A.30}$$

In the sequel, we consider

$$\mathcal{E}_N=\widetilde{\mathcal{E}}_N\cap\left\{\sup_{x\in[0,1]}|F_N^{-1}(x)-F_X^{-1}(x)|\leq N^{-1/3}(\log N)^{-1}\right\}.$$

It follows from (A.27) and (A.30) that $1-\mathbb{P}(\mathcal{E}_N)\ll N^{-\theta}$ so in particular, $\mathbb{P}(\mathcal{E}_N)\geq 1-N^{-\theta}$ for sufficiently large N. It remains to show that (A.29) holds on

$\mathcal{E}_N$. Since $X_1, \ldots, X_N$ are independent with a continuous distribution function, they are all distinct from each other and for all i, there exists a (unique) random j such that $X_i = F_N^{-1}(j/N)$, where F_N^{-1} is the empirical quantile function corresponding to $X_1, \ldots, X_N$. Hence, reordering the terms in the sum in (A.23), we obtain that

$$\begin{aligned} & f(a,u) \\ & \quad = \frac{1}{N}\sum_{i=1}^{N}\left(F_N^{-1}(i/N) - \frac{[Kg(a)]}{K}\right)\left(\mathbb{1}_{F_N^{-1}(iN^{-1})\le u} - \mathbb{1}_{F_N^{-1}(iN^{-1})\le [Kg(a)]K^{-1}}\right) \\ & \quad = \frac{1}{N}\sum_{i=1}^{N}\left(F_N^{-1}(i/N) - \frac{[Kg(a)]}{K}\right)\left(\mathbb{1}_{iN^{-1}\le F_N(u)} - \mathbb{1}_{iN^{-1}\le F_N([Kg(a)]K^{-1})}\right). \end{aligned}$$

Using that F_N^{-1} is constant on all intervals $((i-1)N^{-1}, iN^{-1}]$ we arrive at

$$f(a,u) \quad = \quad \int_{F_N([Kg(a)]K^{-1})}^{F_N(u)}\left(F_N^{-1}(x) - \frac{[Kg(a)]}{K}\right)dx.$$

Hence, on $\mathcal{E}_N$ we have

$$\begin{aligned} & \left|f(a,u) - \int_{F_N([Kg(a)]K^{-1})}^{F_N(u)}\left(F_X^{-1}(x) - \frac{[Kg(a)]}{K}\right)dx\right| \\ & \qquad \le \left|F_N(u) - F_N([Kg(a)]K^{-1})\right| \times \sup_{x\in[0,1]}|F_N^{-1}(x) - F_X^{-1}(x)| \\ & \qquad \le C_2\left(|u-g(a)| + K^{-1} + 2N^{-1/3}(\log N)^{-1}\right)N^{-1/3}(\log N)^{-1} \end{aligned}$$

for all a, u. Hence,

$$\begin{aligned} & \left|f(a,u) - \int_{F_X(g(a))}^{F_X(u)}\big(F_X^{-1}(x) - g(a)\big)dx\right| \\ & \quad \le C_2\big(|u-g(a)| + K^{-1} + 2N^{-1/3}(\log N)^{-1}\big)N^{-1/3}(\log N)^{-1} \\ & \quad + \left|\int_{F_N([Kg(a)]K^{-1})}^{F_N(u)}\left(F_X^{-1}(x) - \frac{[Kg(a)]}{K}\right)dx - \int_{F_X(g(a))}^{F_X(u)}\big(F_X^{-1}(x) - g(a)\big)dx\right| \end{aligned}$$

It follows that

$$\begin{aligned} & \left|f(a,u) - \int_{F_X(g(a))}^{F_X(u)}\big(F_X^{-1}(x) - g(a)\big)dx\right| \\ & \quad \le C_2\big(|u-g(a)| + K^{-1} + 2N^{-1/3}(\log N)^{-1}\big)N^{-1/3}(\log N)^{-1} \\ & \quad + K^{-1}\big|F_X(u) - F_X(g(a))\big| \\ & \quad + \left|\int_{F_N\left(\frac{[Kg(a)]}{K}\right)}^{F_N(u)}\left[F_X^{-1}(x) - \frac{[Kg(a)]}{K}\right]dx - \int_{F_X(g(a))}^{F_X(u)}\left[F_X^{-1}(x) - \frac{[Kg(a)]}{K}\right]dx\right|. \end{aligned}$$

Now, on $\mathcal{E}_N$ we also have

$$\begin{aligned}\left|F_X^{-1}(x)-\frac{[Kg(a)]}{K}\right| &= \left|F_X^{-1}(x)-F_X^{-1}\circ F_X\left(\frac{[Kg(a)]}{K}\right)\right| \\ &\leq \frac{1}{C_1}\left|x-F_X\left(\frac{[Kg(a)]}{K}\right)\right| \\ &\leq \frac{1}{C_1}\left(\left|F_N(u)-F_N\left(\frac{[Kg(a)]}{K}\right)\right|+C_2N^{-1/3}(\log N)^{-1}\right),\end{aligned}$$

for all x lying between $F_N(u)$ and $F_N([Kg(a)]K^{-1})$. For such x's, we obtain on $\mathcal{E}_N$ that

$$\begin{aligned}\left|F_X^{-1}(x)-\frac{[Kg(a)]}{K}\right| &\leq \frac{1}{C_1}\left(\left|F_X(u)-F_X\left(\frac{[Kg(a)]}{K}\right)\right|+3C_2N^{-1/3}(\log N)^{-1}\right) \\ &\leq \frac{3C_2}{C_1}\left(|u-g(a)|+K^{-1}+N^{-1/3}(\log N)^{-1}\right)\end{aligned}$$

for all a and u, for sufficiently large N. Therefore, with $K\geq 1$ we obtain on $\mathcal{E}_N$ that

$$\begin{aligned}&\left|f(a,u)-\int_{F_X(g(a))}^{F_X(u)}\left(F_X^{-1}(x)-g(a)\right)dx\right| \\ &\quad\leq 2C_2\left(|u-g(a)|+K^{-1}+2N^{-1/3}(\log N)^{-1}\right)N^{-1/3}(\log N)^{-1} \\ &\qquad+\frac{3C_2}{C_1}\left(|u-g(a)|+K^{-1}+N^{-1/3}(\log N)^{-1}\right) \\ &\qquad\times\left(2\sup_{u\in[0,1]}|F_N(u)-F_X(u)|+C_2K^{-1})\right) \\ &\quad= O\left(|u-g(a)|+K^{-1}+N^{-1/3}(\log N)^{-1}\right)\left(N^{-1/3}(\log N)^{-1}+K^{-1}\right)\end{aligned}$$

on $\mathcal{E}_N$, uniformly over $a\in\mathbb{R}$ and $u\in\{\overline{x}_0,\ldots,\overline{x}_K\}$. Now, we can do the change of variable $t=F_X^{-1}(x)$ to get on $\mathcal{E}_N$ that

$$\begin{aligned}&f(a,u) \\ &\quad=\int_{g(a)}^{u}(t-g(a))\,f_X(t)dt \qquad\qquad (A.31)\\ &\quad+O\left(|u-g(a)|+K^{-1}+N^{-1/3}(\log N)^{-1}\right)\left(N^{-1/3}(\log N)^{-1}+K^{-1}\right)\end{aligned}$$

uniformly over $a\in\mathbb{R}$ and $u\in\{\overline{x}_0,\ldots,\overline{x}_K\}$. Here,

$$\int_{g(a)}^{u}(t-g(a))\,f_X(t)dt \quad\geq\quad C_1\int_{g(a)}^{u}(t-g(a))\,dt$$

where C_1 is taken from (3.2), for all a,u. Since it is assumed that $K^{-1}=o(N^{-1/3})$, we conclude that on $\mathcal{E}_N$,

$$f(a,u) \quad\geq\quad \frac{C_1}{2}(u-g(a))^2+o((g(a)-u)^2),$$

where the small o-term is uniform over all u and a such that $|u-g(a)| \geq N^{-1/3}$.

Hence, (A.29) holds on $\mathcal{E}_N$ provided that $\tilde{c} < C_1/2$ and N is sufficiently large. It follows that on $\mathcal{E}_N$, for all sufficiently large N,

$$e(a,u) \leq o((u-g(a))^2 - C_3\tilde{c}(g(a)-u)^2$$

where in view of the above proof, the small-o term can be chosen of the form

$$o((u-g(a))^2 = 2C_2C_4K^{-1}(|u-g(a)| + N^{-1/3}).$$

Therefore, for any $c < C_3\tilde{c}$, for all sufficiently large N, $e(a,u) \leq -c(g(a)-u)^2$ on $\mathcal{E}_N$. This completes the proof of the lemma. □

A.5. Proof of Lemma 7.2

The proof rests on the following proposition.

Proposition P. Let $\mathcal{F}$ be a class of functions from $\mathcal{X}$ to $\mathbb{R}$ and let $X_1, X_2, \ldots, X_n$ be independent (but not necessarily identically distributed) random variables defined on $\mathcal{X}$. Let F be a measurable envelope for the class $\mathcal{F}$ and assume that $E(F^2(X_i)) < \infty$ for each $1 \leq i \leq n$. Define

$$\mathbb{G}_n f := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (f(X_i) - Ef(X_i))$$

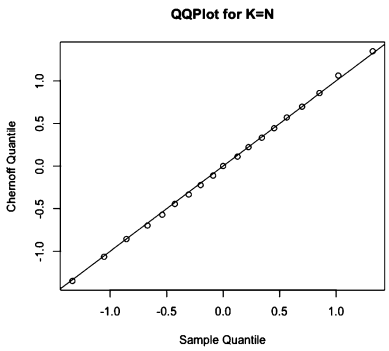
and let $\|\mathbb{G}_n\|_{\mathcal{F}} := \sup_{f\in\mathcal{F}} |\mathbb{G}_n f|$. Then, for $p \geq 2$, there exists $A_p > 0$ that depends on p only such that

$$E^*\left[\|\mathbb{G}_n\|_{\mathcal{F}}^p\right] \leq A_p \mathcal{J}(1,\mathcal{F}) \frac{1}{n} \sum_{i=1}^{n} E(F^p(X_i)),$$

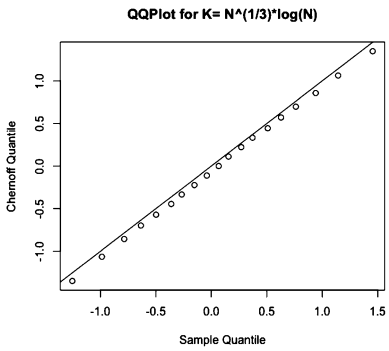
where E^* denotes outer expectation and $\mathcal{J}(1,\mathcal{F})$ is taken from Section 2.14.1 of [23].

The proof of the Proposition P is essentially the same as the proof of Theorem 2.14.1 in [23], by noting that the steps in the proof remain valid even if the X_i's are not i.i.d but are independent with potentially different distributions. Hence, the proof is omitted.

We now apply the above proposition with n replaced by N, $\mathcal{F} := \{f_w(\cdot) : |w| \leq x\}$ where $f_w(t) = 1(t \leq u_0 + w) - 1(u \leq u_0)$ with envelope function $F(t) = 1(t \leq u_0 + x) - 1(t \leq u_0 - x)$, and $p = 2$. The supremum in $\|\mathbb{G}_N\|_{\mathcal{F}}$ is taken over a finite set and hence, it is measurable. This implies that the outer expectation can be replaced by an expectation in Proposition P. Moreover, the class of functions under consideration is a VC class of dimension 3 and therefore $\mathcal{J}(1,\mathcal{F}) < \infty$. Note that $E(F^2(X_i)) \leq C_2 x$ since the densities of the X_i's satisfy (4.2). The assertion of the lemma now follows directly from Proposition P upon dividing both sides of the inequality in the proposition by N. □



(a) QQPlot: Asymptotic distribution versus Global Estimator



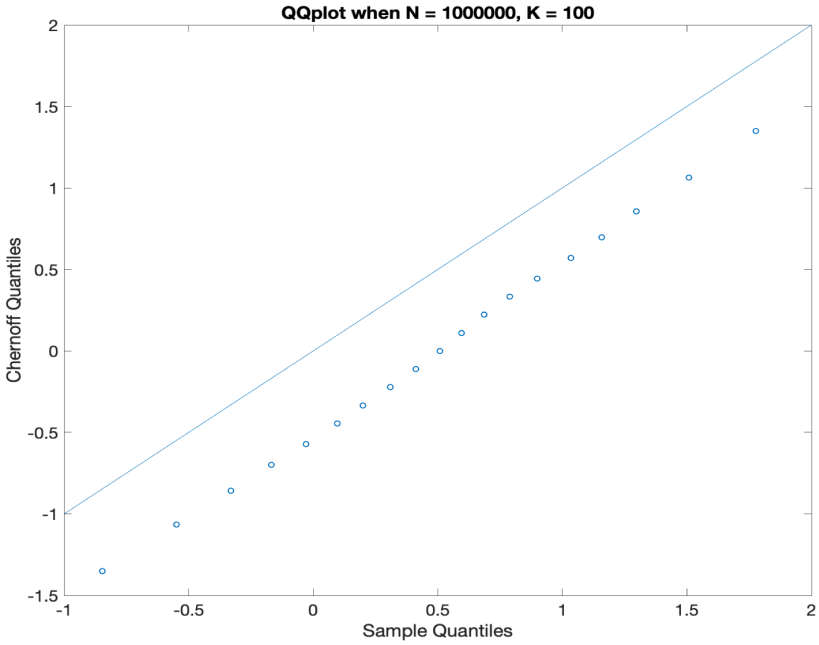
(b) QQPlot: Asymptotic distribution versus Pooled Estimator

A.6. Limited simulation results

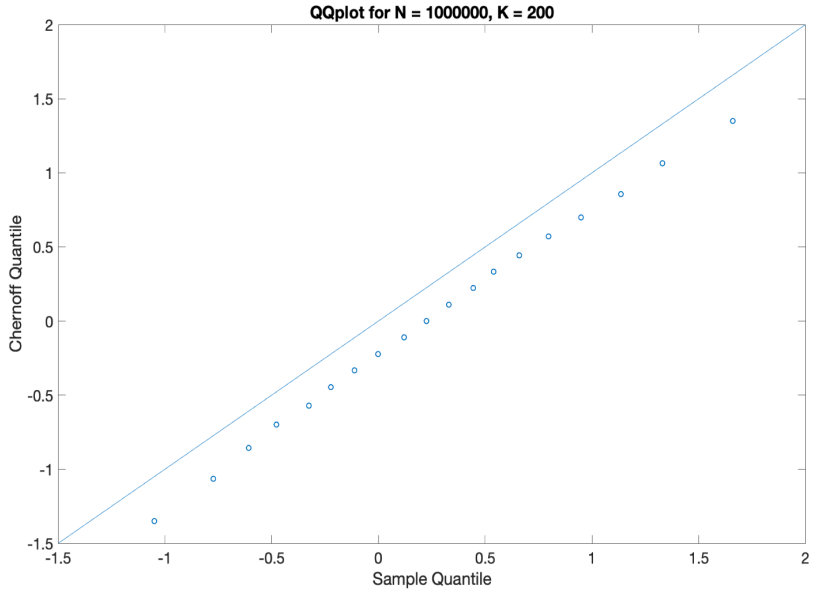
Simulation results are presented for the following setting. The model is $Y = X^2 + \epsilon$ with $X \sim \text{Unif}(0,1)$ and $\epsilon \sim N(0,1)$ independently of X. We are interested in estimating $\mu(x) \equiv x^2$ at the point $x_0 = 0.5$. The limit distribution of the global isotonic estimator, i.e. the limiting law of $N^{1/3}(\hat{\mu}_G(x_0) - \mu(x_0))$ is given, for example, in Equation (1.2) of BDSE. Note that the setting considered in the simulations corresponds to the homogeneous case, i.e. the case where $m = 1$.

We took $N = 10^6$ and generated 5000 replicates from the distribution of the global isotonic estimator, and the isotonic estimator formed by the binning procedure with $K = N^{1/3} \log N$. Selected quantiles from the empirical distributions of $N^{1/3}(\hat{\mu}_G(x_0) - \mu(x_0))$ and $N^{1/3}(\hat{\mu}_{binned}(x_0) - \mu(x_0))$ [based on the 5000 replicates] were then compared to the quantiles of the limit distribution which were generated by plugging in the true value of κ that appears in (1.2) of BDSE and using the numerical values of the quantiles of Chernoff's distribution provided in [9]. The QQ-plots are given in the left and right panels of the first figure respectively: both panels indicate close agreement of the empirical distributions with the truth. In general, the prescription $K = N^{1/3} \log N$ does quite well, when N is in the order of millions or larger. Smaller N's are not terribly interesting from the big data perspective and were not considered. For smaller N, like in the thousands, much more fine-tuning will be needed to find an adequate K but this is not too relevant to the problem this paper seeks to address.

The second figure shows that with binning of order $N^{1/3}$ the limit distribution deviates from the Chernoff limit. The second and third figures show the QQ-plots when we use $K = N^{1/3}$ and $K = 2N^{1/3}$ respectively. In both cases, the empirical distribution deviates systematically from the limit, with the deviation in the former case much more pronounced owing to a larger bias.



(a) Binned estimator with $K = N^{1/3}$



(b) Binned estimator with $K = 2N^{1/3}$

References

[1] Anevski, D., Hössjer, O., et al. (2006). A general asymptotic scheme for inference under order restrictions. *The Annals of Statistics*, 34(4):1874–1930. MR2283721

[2] Azadbakhsh, M., Jankowski, H., and Gao, X. (2014). Computing confidence intervals for log-concave densities. *Computational Statistics & Data Analysis*, 75:248–264. MR3178372

[3] Banerjee, M., Durot, C., Sen, B., et al. (2019). Divide and conquer in nonstandard problems and the super-efficiency phenomenon. *The Annals of Statistics*, 47(2):720–757. MR3909948

[4] Banerjee, M. et al. (2007). Likelihood based inference for monotone response models. *The Annals of Statistics*, 35(3):931–956. MR2341693

[5] Banerjee, M. and Wellner, J. A. (2001). Likelihood ratio tests for monotone functions. *Annals of Statistics*, pages 1699–1731. MR1891743

[6] Battey, H., Fan, J., Liu, H., Lu, J., and Zhu, Z. (2018). Distributed testing and estimation under sparse high dimensional models. *Annals of statistics*, 46(3):1352. MR3798006

[7] Billingsley, P. (2013). *Convergence of probability measures*. John Wiley & Sons. MR0233396

[8] Groeneboom, P., Jongbloed, G., Witte, B. I., et al. (2010). Maximum smoothed likelihood estimation and smoothed maximum likelihood estimation in the current status model. *The Annals of Statistics*, 38(1):352–387. MR2589325

[9] Groeneboom, P. and Wellner, J. A. (2001). Computing Chernoff's distribution. *J. Comput. Graph. Statist.*, 10(2):388–400. MR1939706

[10] Hsieh, C.-J., Si, S., and Dhillon, I. (2014). A divide-and-conquer solver for kernel support vector machines. In *International Conference on Machine Learning*, pages 566–574.

[11] Kim, J. and Pollard, D. (1990). Cube root asymptotics. *Annals of Statistics*, 18:191–219. MR1041391

[12] 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. MR3117826

[13] Lu, J., Cheng, G., and Liu, H. (2016). Nonparametric heterogeneity testing for massive data. *arXiv preprint* arXiv:1601.06212.

[14] Mammen, E. (1991). Estimating a smooth monotone regression function. *The Annals of Statistics*, pages 724–740. MR1105841

[15] Massart, P. (1990). The tight constant in the dvoretzky-kiefer-wolfowitz inequality. *The Annals of Probability*, pages 1269–1283. MR1062069

[16] Mukerjee, H. (1988). Monotone nonparametric regression. *The Annals of Statistics*, pages 741–750. MR0947574

[17] 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. MR0961262

[18] Rosenthal, H. P. (1970). On the subspaces of $L_p, (p > 2)$ spanned by sequences of independent random variables. *Israel Journal of Mathematics*, 8(3):273–303. MR0271721

[19] Shang, Z. and Cheng, G. (2017). Computational limits of a distributed algorithm for smoothing spline. *The Journal of Machine Learning Research*, 18(1):3809–3845. MR3725447

[20] Shi, C., Lu, W., and Song, R. (2018). A massive data framework for m-estimators with cubic-rate. *Journal of the American Statistical Association*, 113(524):1698–1709. MR3902239

[21] Tang, R., Banerjee, M., Kosorok, M. R., et al. (2012). Likelihood based inference for current status data on a grid: A boundary phenomenon and an adaptive inference procedure. *The Annals of Statistics*, 40(1):45–72. MR3013179

[22] Van Der Vaart, A. and Van Der Laan, M. (2003). Smooth estimation of a monotone density. *Statistics: A Journal of Theoretical and Applied Statistics*, 37(3):189–203. MR1986176

[23] van der Vaart, A. W. and Wellner, J. A. (1996). *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York. With applications to statistics. MR1385671

[24] Volgushev, S., Chao, S.-K., and Cheng, G. (2019). Distributed inference for quantile regression processes. *The Annals of Statistics*, 47(3):1634–1662. MR3911125

[25] Zhang, R., Kim, J., and Woodroofe, M. (2001). Asymptotic analysis of isotonic estimation for grouped data. *Journal of statistical planning and inference*, 98(1–2):107–117. MR1860229

[26] Zhang, Y., Duchi, J., and Wainwright, M. (2013). Divide and conquer kernel ridge regression. In *Conference on Learning Theory*, pages 592–617. MR3450540

[27] Zhao, T., Cheng, G., and Liu, H. (2016). A partially linear framework for massive heterogeneous data. *Annals of statistics*, 44(4):1400. MR3519928