Journal of Machine Learning Research 23 (2022) 1-61 Submitted 10/21; Revised 6/22; Published 8/22

Communication-Constrained Distributed Quantile Regression with Optimal Statistical Guarantees

Kean Ming Tan KEANMING@UMICH.EDU
Department of Statistics
University of Michigan
Ann Arbor MI, 48109, USA

Heather Battey H.BATTEY@IMPERIAL.AC.UK
Department of Mathematics
Imperial College London
London, SW7 2AZ, U.K.

Wen-Xin Zhou WEZ243@UCSD.EDU
Department of Mathematical Sciences
University of California, San Diego
La Jolla, CA 92093, USA

Editor: Qiang Liu

Abstract

We address the problem of how to achieve optimal inference in distributed quantile regression without stringent scaling conditions. This is challenging due to the non-smooth nature of the quantile regression (QR) loss function, which invalidates the use of existing methodology. The difficulties are resolved through a double-smoothing approach that is applied to the local (at each data source) and global objective functions. Despite the reliance on a delicate combination of local and global smoothing parameters, the quantile regression model is fully parametric, thereby facilitating interpretation. In the low-dimensional regime, we establish a finite-sample theoretical framework for the sequentially defined distributed QR estimators. This reveals a trade-off between the communication cost and statistical error. We further discuss and compare several alternative confidence set constructions, based on inversion of Wald and score-type tests and resampling techniques, detailing an improvement that is effective for more extreme quantile coefficients. In high dimensions, a sparse framework is adopted, where the proposed doubly-smoothed objective function is complemented with an ℓ_1-penalty. We show that the corresponding distributed penalized QR estimator achieves the global convergence rate after a near-constant number of communication rounds. A thorough simulation study further elucidates our findings.

Keywords: Communication efficiency; convolution smoothing; data heterogeneity; decentralized learning; distributed inference; multiplier bootstrap; quantile regression.

1. Introduction

Quantile regression is indispensable for understanding pathways of dependence irretrievable through a standard conditional mean regression analysis. Since its inception by Koenker and Bassett (1978), appreciable effort has been expended in understanding and operationalizing quantile regression. Statistical aspects have focused on the situation in which all the

data are simultaneously available for inference while practical aspects have centered around reformulations of the quantile regression optimization problem for computational efficiency.

Challenges arise when data are distributed, either by the study design or due to storage and privacy concerns. The latter have become more prominent, with less centralized systems tending to be preferred both by the individuals whose data are collected and by those responsible for ensuring their security. In such settings, communication costs associated with statistical procedures become a consideration in addition to their theoretical properties. Ideally, inferential tools are sought whose communication costs are as low as possible without sacrificing statistical accuracy, where the latter would be quantified in terms of estimation error or distributional approximation errors for test statistics.

Data may be naturally partitioned because of the way they were collected, or deliberately distributed for other reasons. Li *et al.* (2020) provided examples in which the distributed setting arises: (i) when data are too numerous to be stored in a central location; (ii) when privacy and security are a concern, such as for medical records, necessitating decentralized statistical analyses. We are motivated particularly by situations in which there are separate data-collecting entities such as local governments, research labs, hospitals, or smart phones, and direct data sharing raises concerns over privacy or loss of ownership. Due to privacy concerns over sending raw data, data collected at each location must remain there, which makes communication efficiency critical, especially when the network comprises an enormous number of local data-collecting entities. Communication in the network can be slower than local computation by three orders of magnitude due to limited bandwidth (Lan *et al.*, 2020). It is therefore desirable to communicate as few rounds as possible, leaving expensive computation to local machines.

Among two general principles that have been proposed for distributed statistical inference, the simple meta-analysis approach of averaging estimates from separate data sources has the advantage of only requiring one round of communication. Jordan, Lee and Yang (2019) highlighted some disadvantages. Notably, in a simpler setting than that posed by quantile regression, a stringent constraint on the number of sources is implicated. To attain the convergence rate hypothetically achievable using the combined sample of size N, a meta-analysis must limit the number of sources, m, to be far fewer than $\sqrt{N}$. This is due to small sample bias inherent to most nonlinear estimators, which does not diminish upon aggregation. A violation of the scaling condition slows the convergence rate of the estimator. This, while sometimes acceptable for point estimation, is detrimental for statistical inference, as illustrated later in simulations.

By extending the distributed approximate Newton algorithm (Shamir, Srebro and Zhang, 2014), Wang *et al.* (2017) and Jordan, Lee and Yang (2019) proposed an alternative principle for distributed inference in parametric models, which requires a controlled amount of further communications to yield statistically optimal estimators without the restriction $m = o(\sqrt{N})$ on the number of machines. Another variant of this principle was considered in Fan, Guo and Wang (2021), along with a simultaneous analysis of the optimization and statistical errors. For reasons outlined below, these ideas are not directly applicable to quantile regression without considerable methodological development, guided by the detailed theoretical analysis provided in this paper.

Quantile regression quantifies dependence of an outcome variable's quantiles on a number of covariates. All quantiles are potentially of interest but to take an important example,

quantile regression has bearing on the types of applications for which conditional extreme value analysis might be considered. There are relatively few successful examples of modeling the extremes. The pioneering work of Engelke and Hitz (2020) being a notable exception, suitable when all variables are on an equal footing. When explanation for extreme behavior of a particular variable is sought, as would be the case in many hydrological, sociological, and medical applications, quantile regression provides succinct interpretable conclusions. Subtle graphical structure is deducible from a succession of quantile regression analyses by a result of Cox (2007), generalizing insights by Cochran (1938). Besides substantive understanding furnished by a quantile regression model, coefficient estimators enjoy robustness properties in the form of limited sensitivity to anomalous data or leptokurtic tail behavior of the conditional distribution of the outcome.

The associated non-differentiable loss function, otherwise responsible for tortuously slow computation, necessitates linear programming reformulations, solvable by variants of simplex and interior point methods. These algorithms are not compatible with distributed architectures, rendering statistical inference challenging when data are distributed. Even when computation is ignored and the non-differentiable loss function is used directly, both the distributed estimation procedures proposed by Wang *et al.* (2017), Jordan, Lee and Yang (2019) and Fan, Guo and Wang (2021) and the technical devices used therein are unavailable due to their requirements on the loss function. Namely that it be strongly convex and twice differentiable with Lipschitz continuous second derivatives.

Two papers by Volgushev, Chao and Cheng (2019) and Chen, Liu and Zhang (2021) are motivated by the challenges of distributed data and the relevance of quantile regression, seeking synthesized estimators of quantile regression coefficients. These papers employed the meta-analysis approach, thereby requiring stringent scaling to achieve the desired theoretical guarantees, although with the advantage of requiring a single round of communication. We discuss these works in greater detail in Section 2.4. In addition to the scaling deficiencies, a generalization of the simple meta-analysis approach to high-dimensional settings has proved elusive for quantile regression. In sparse high-dimensional linear and generalized linear models, the success of meta-analyses hinges on the ability to de-bias suitably penalized estimators (Lee *et al.*, 2017; Battey *et al.*, 2018). Such de-biased estimators are unavailable for penalized quantile regression, except under the stringent assumption that the regression error is independent of the covariates (Bradic and Kolar, 2017).

The present paper operationalizes the ideas of Jordan, Lee and Yang (2019) and Wang *et al.* (2017) in the context of quantile regression, enabling distributed estimation and inference in low and high-dimensional regimes. The key idea of our proposal is double-smoothing of the local and global approximate loss functions, which requires different smoothing bandwidths to achieve desirable statistical properties. Specifically, our proposed synthesized estimator achieves the optimal statistical rate of convergence by a delicate combination of local and global smoothing, and number of communication rounds. The latter turns out to be small.

In the low-dimensional regime, we further detail distributed constructions of confidence sets. Among these is one based on a self-normalized reformulation of a score-type statistic. Modulo estimation of the parameter vector, score constructions rewritten as self-normalized sums enjoy a form of linearity that enables synthesis across data sources without information loss. To our knowledge, this work is the first to provide Berry-Esseen type quantification of

distributional approximation errors in a distributed setting, which may be of independent interest. In the high-dimensional regime, the proposed doubly-smoothed local and global objective functions are coupled with an ℓ_1 penalty to encourage sparse solutions, which we solve using a locally adaptive majorize-minimize algorithm. Theoretically, we show that the resulting estimator is near-optimal under both the ℓ_1 and ℓ_2 norms. The results are presented in Section 3.

Notation: For every integer $k \geq 1$, we use $\mathbb{R}^k$ to denote the the k-dimensional Euclidean space. The inner product of any two vectors $\boldsymbol{u} = (u_1, \ldots, u_k)^{\mathrm{T}}, \boldsymbol{v} = (v_1, \ldots, v_k)^{\mathrm{T}} \in \mathbb{R}^k$ is defined by $\boldsymbol{u}^{\mathrm{T}}\boldsymbol{v} = \langle \boldsymbol{u}, \boldsymbol{v} \rangle = \sum_{i=1}^k u_i v_i$. We use $\|\cdot\|_p$ $(1 \leq p \leq \infty)$ to denote the ℓ_p-norm in $\mathbb{R}^k$: $\|\boldsymbol{u}\|_p = (\sum_{i=1}^k |u_i|^p)^{1/p}$ and $\|\boldsymbol{u}\|_\infty = \max_{1 \leq i \leq k} |u_i|$. Throughout this paper, we use bold capital letters to represent matrices. For $k \geq 2$, $\mathbf{I}_k$ represents the identity matrix of size k. For any $k \times k$ symmetric matrix $\mathbf{A} \in \mathbb{R}^{k \times k}$, $\|\mathbf{A}\|_2$ is the operator norm of $\mathbf{A}$. For a positive semidefinite matrix $\mathbf{A} \in \mathbb{R}^{k \times k}$, $\|\cdot\|_{\mathbf{A}}$ denotes the norm linked to $\mathbf{A}$ given by $\|\boldsymbol{u}\|_{\mathbf{A}} = \|\mathbf{A}^{1/2}\boldsymbol{u}\|_2$, $\boldsymbol{u} \in \mathbb{R}^k$. Moreover, given $r \geq 0$, define the Euclidean ball and sphere in $\mathbb{R}^k$ as $\mathbb{B}^k(r) = \{\boldsymbol{u} \in \mathbb{R}^k : \|\boldsymbol{u}\|_2 \leq r\}$ and $\mathbb{S}^{k-1}(r) = \partial\mathbb{B}^k(r) = \{\boldsymbol{u} \in \mathbb{R}^k : \|\boldsymbol{u}\|_2 = r\}$, respectively. In particular, $\mathbb{S}^{k-1} \equiv \mathbb{S}^{k-1}(1)$ denotes the unit sphere. For two sequences of non-negative numbers $\{a_n\}_{n \geq 1}$ and $\{b_n\}_{n \geq 1}$, $a_n \lesssim b_n$ indicates that there exists a constant $C > 0$ independent of n such that $a_n \leq C b_n$; $a_n \gtrsim b_n$ is equivalent to $b_n \lesssim a_n$; $a_n \asymp b_n$ is equivalent to $a_n \lesssim b_n$ and $b_n \lesssim a_n$.

2. Distributed Inference for Quantile Regression

2.1 Conquer: convolution-smoothed quantile regression

For a quantile level $\tau \in (0, 1)$, we consider a linear conditional quantile model for the data vector $(y, \boldsymbol{x}) \in \mathbb{R} \times \mathbb{R}^p$:

$$Q_\tau(y|\boldsymbol{x}) = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^* = \sum_{j=1}^p x_j \beta_j^*, \tag{1}$$

where $Q_\tau(y|\boldsymbol{x})$ denotes the conditional τ-quantile of y given $\boldsymbol{x} = (x_1, \ldots, x_p)^{\mathrm{T}}$ with $x_1 \equiv 1$. Here, $\boldsymbol{\beta}^* = \boldsymbol{\beta}^*(\tau) \in \mathbb{R}^p$ is the regression coefficient vector that minimizes the criterion function

$$\mathcal{Q}(\boldsymbol{\beta}) := \mathbb{E}\big\{\rho_\tau(y - \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta})\big\}, \tag{2}$$

where $\rho_\tau(u) = u\{\tau - \mathbb{1}(u < 0)\}$ is the asymmetric absolute deviation function, also known as the *check function* or *pinball loss*. Given a random sample $\{(y_i, \boldsymbol{x}_i)\}_{i=1}^N$ of size $N > p$ from $(y, \boldsymbol{x})$, the linear quantile regression estimator of $\boldsymbol{\beta}^*$ is defined as a minimizer of the empirical analog of $\mathcal{Q}(\cdot)$:

$$\widehat{\boldsymbol{\beta}} \in \operatorname*{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^p} \widehat{\mathcal{Q}}(\boldsymbol{\beta}), \quad \text{where} \quad \widehat{\mathcal{Q}}(\boldsymbol{\beta}) := \frac{1}{N}\sum_{i=1}^N \rho_\tau(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}). \tag{3}$$

Since the seminal work of Koenker and Bassett (1978), quantile regression (QR) has been extensively studied from both statistical and computational perspectives. We refer to Koenker (2005) and Koenker *et al.* (2017) for a systematic introduction of quantile regression under various settings.

By the convexity of the check function, the population loss function $\mathcal{Q}(\cdot)$ in (2) is also convex. Moreover, under mild conditions, $\mathcal{Q}(\cdot)$ is twice differentiable and strongly convex in a neighborhood of $\boldsymbol{\beta}^*$ with Hessian matrix $\mathbf{H} := \nabla^2\mathcal{Q}(\boldsymbol{\beta}^*) = \mathbb{E}\{f_{\varepsilon|\boldsymbol{x}}(0)\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\}$, where $f_{\varepsilon|\boldsymbol{x}}(\cdot)$ denotes the conditional density of ε given $\boldsymbol{x}$. In contrast, the empirical loss $\widehat{\mathcal{Q}}(\cdot)$ is not differentiable at $\boldsymbol{\beta}^*$, and its "curvature energy" is concentrated at a single point. This is substantially different from other widely used loss functions that are at least locally strongly convex, such as the squared or logistic loss. To deal with the non-smoothness issue, Horowitz (1998) proposed to smooth the objective function, or equivalently the check function $\rho_\tau(\cdot)$, to obtain $\rho_\tau^{\mathrm{H}}(u) = u\{\tau - G(-u/h)\}$, where $G(\cdot)$ is a smooth function and $h > 0$ is the smoothing parameter or bandwidth. See also Wang, Stefanski and Zhu (2012), Wu, Ma and Yin (2015), Galvao and Kato (2016) and Chen, Liu and Zhang (2019) for extensions of such a smoothed objective function approach with more complex data. However, Horowitz's smoothing gains smoothness at the cost of convexity, which inevitably raises optimization issues especially when p is large. On the other hand, by the first-order condition, the population parameter $\boldsymbol{\beta}^*$ satisfies the moment condition $\nabla\mathcal{Q}(\boldsymbol{\beta}^*) = \mathbb{E}[\{\mathbb{1}(y < \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}) - \tau\}\boldsymbol{x}]|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*} = \mathbf{0}$. This property motivates a smoothed estimating equation (SEE) estimator (Whang, 2006; Kaplan and Sun, 2017), defined as the solution to the smoothed moment condition

$$\frac{1}{N}\sum_{i=1}^{N}\big\{G\big((\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} - y_i)/h\big) - \tau\big\}\boldsymbol{x}_i = \mathbf{0}. \tag{4}$$

From an M-estimation viewpoint, the aforementioned SEE estimator can be equivalently defined as a minimizer of the empirical smoothed loss function

$$\widehat{\mathcal{Q}}_h(\boldsymbol{\beta}) = \frac{1}{N}\sum_{i=1}^{N}\ell_h(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}) \quad \text{with} \quad \ell_h(u) = (\rho_\tau * K_h)(u) = \int_{-\infty}^{\infty}\rho_\tau(v)K_h(v-u)\,\mathrm{d}v, \tag{5}$$

where $K(\cdot)$ is a kernel function, $K_h(u) = (1/h)K(u/h)$, and $*$ is the convolution operator. This approach will be referred to as *conquer*, which stands for convolution-type smoothed quantile regression. The ensuing estimator is then denoted by $\widehat{\boldsymbol{\beta}}^{\mathrm{cq}} = \widehat{\boldsymbol{\beta}}_h^{\mathrm{cq}} \in \operatorname{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p}\widehat{\mathcal{Q}}_h(\boldsymbol{\beta})$.

To see the connection between SEE and conquer methods, define $\overline{K}(u) = \int_{-\infty}^{u}K(v)\,\mathrm{d}v$, and note that the empirical loss $\widehat{\mathcal{Q}}_h(\cdot)$ in (5) is twice continuously differentiable with gradient and Hessian given by $\nabla\widehat{\mathcal{Q}}_h(\boldsymbol{\beta}) = (1/N)\sum_{i=1}^{N}\{\overline{K}((\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} - y_i)/h) - \tau\}\boldsymbol{x}_i$ and $\nabla^2\widehat{\mathcal{Q}}_h(\boldsymbol{\beta}) = (1/N)\sum_{i=1}^{N}K_h(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})\cdot\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}$, respectively. When a non-negative kernel is used, $\widehat{\mathcal{Q}}_h(\cdot)$ is convex so that any minimizer of $\boldsymbol{\beta}\mapsto\widehat{\mathcal{Q}}_h(\boldsymbol{\beta})$ satisfies the first-order moment condition (4) with $G = \overline{K}$.

When the dimension p is fixed, asymptotic properties of the SEE or conquer estimator have been studied by Kaplan and Sun (2017) and Fernandes, Guerre and Horta (2021), although the former focused on a more challenging instrumental variables quantile regression problem. In the finite sample setup, He *et al.* (2021) established exponential-type concentration inequalities and nonasymptotic Bahadur representation for the conquer estimator, while allowing the dimension p to grow with the sample size n. Their results reveal a key feature of the smoothing parameter: the bandwidth should adapt to both the sample size n

and dimensionality p, so as to achieve a trade-off between statistical accuracy and computational stability. For statistical inference, He *et al.* (2021) suggested and proved the validity of the *multiplier bootstrap* for conquer, which has desirable finite sample performance under various settings, including those at extreme quantile levels. We refer to Section 5 of He *et al.* (2021) for further details on the computational aspects of conquer.

2.2 Distributed quantile regression with conquer

Before detailing an approach for distributed inference for QR coefficients, motivated primarily by situations in which the data are distributed, we start with some remarks on computation. The optimization problem in (3) can be recast as a convex linear program, solvable by the simplex or interior point methods. The latter has a computational complexity of order $\mathcal{O}(N^{1+a}p^3 \log N)$ for some $a \in (0, 1/2)$. An efficient algorithm, the Frisch-Newton algorithm with preprocessing, has an improved complexity of $\mathcal{O}\{(Np)^{2(1+a)/3}p^3 \log N + Np\}$ (Portnoy and Koenker, 1997). While not inordinate relative to the $\mathcal{O}(p^2N)$ complexity of least squares, to achieve the same quality of distributional approximation, quantile regression requires a considerably larger sample size. Thus for formal inference there are sometimes computational advantages to parallelized inference even when data are available in their totality.

For ease of exposition, assume that the m data sources are of equal sample size n, so that $N = m \cdot n$. The combined data set is $\{(y_i, \boldsymbol{x}_i)\}_{i=1}^N$, where $\boldsymbol{x}_i$ is a p-dimensional vector. For $j = 1, \ldots, m$, the jth location stores a subsample of n observations, denoted by $\mathcal{D}_j = \{(y_i, \boldsymbol{x}_i)\}_{i \in \mathcal{I}_j}$, and $\{\mathcal{I}_j\}_{j=1}^m$ are disjoint index sets satisfying $\cup_{j=1}^m \mathcal{I}_j = \{1, \ldots, N\}$ and $|\mathcal{I}_j| = n$, where $|\mathcal{I}_j|$ is the cardinality of $\mathcal{I}_j$.

Under a conditional quantile regression model, the observations $(y_1, \boldsymbol{x}_1), \ldots, (y_N, \boldsymbol{x}_N)$ are i.i.d. sampled from $(y, \boldsymbol{x}) \sim P$ satisfying $Q_\tau(y|\boldsymbol{x}) = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*$, and the model parameter $\boldsymbol{\beta}^*$ is equivalently defined as

$$\boldsymbol{\beta}^* = \operatorname*{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^p} \mathcal{Q}(\boldsymbol{\beta}), \quad \mathcal{Q}(\boldsymbol{\beta}) := \mathbb{E}_{(y,\boldsymbol{x}) \sim P}\big\{\rho_\tau\big(y - \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}\big)\big\}, \tag{6}$$

where $\rho_\tau(\cdot)$ is the check function. Unlike the model setting considered by Wang *et al.* (2017) and Jordan, Lee and Yang (2019) in which the target loss function is twice differentiable and has Lipschitz continuous second derivative, the non-smooth check function is not everywhere differentiable, which prevents gradient-based optimization methods from being efficient.

Given two bandwidths $h, b > 0$, we define the global and local smoothed quantile loss functions as

$$\widehat{\mathcal{Q}}_h(\boldsymbol{\beta}) = \frac{1}{N}\sum_{i=1}^N \ell_h(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}) \quad \text{and} \quad \widehat{\mathcal{Q}}_{j,b}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i \in \mathcal{I}_j} \ell_b(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}), \quad j = 1, \ldots, m, \tag{7}$$

where the loss function $\ell_h(\cdot)$ is as defined in (5). Hereafter, h and b will be referred to as the *global bandwidth* and *local bandwidth*, respectively, and we assume $b \geq h > 0$. In the context of quantile regression, we extend the approximate Newton-type method proposed by Shamir, Srebro and Zhang (2014) through convolution smoothing; see also Wang *et al.* (2017) and Jordan, Lee and Yang (2019). Notably, the ideas behind these Newton-type

methods coincide, to some extent, with the classical one-step construction (Bickel, 1975), which focused on improving an initial estimator that is already consistent but not efficient.

Starting with an initial estimator $\widetilde{\boldsymbol{\beta}}^{(0)}$ of $\boldsymbol{\beta}^*$, we define the shifted conquer loss function

$$\widetilde{\mathcal{Q}}(\boldsymbol{\beta}) = \widehat{\mathcal{Q}}_{1,b}(\boldsymbol{\beta}) - \big\langle \nabla \widehat{\mathcal{Q}}_{1,b}(\widetilde{\boldsymbol{\beta}}^{(0)}) - \nabla \widehat{\mathcal{Q}}_h(\widetilde{\boldsymbol{\beta}}^{(0)}), \boldsymbol{\beta} \big\rangle, \tag{8}$$

which leverages local higher-order information and global first-order information, and therefore depends on both local and global bandwidths b and h. The resulting communication-efficient estimator is given by

$$\widetilde{\boldsymbol{\beta}}^{(1)} = \widetilde{\boldsymbol{\beta}}^{(1)}_{b,h} \in \operatorname*{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^p} \widetilde{\mathcal{Q}}(\boldsymbol{\beta}). \tag{9}$$

Informal motivation for the aforementioned approach is provided by a Taylor series expansion of the global loss function around the initial estimator $\widetilde{\boldsymbol{\beta}}^{(0)}$. For a suitable choice of $\widetilde{\boldsymbol{\beta}}^{(0)}$, the approximation error is well controlled in view of the heuristic argument outlined by Jordan, Lee and Yang (2019). Furthermore, on writing $\widetilde{Q}(\boldsymbol{\beta})$ more explicitly as $\widetilde{Q}(\boldsymbol{\beta}; \widetilde{\boldsymbol{\beta}}^{(0)})$, it can be arranged, through bandwidths b and h, that $\operatorname{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^p} \widetilde{Q}(\boldsymbol{\beta}; \widetilde{\boldsymbol{\beta}}^{(0)})$ is a contraction mapping in a suitable neighborhood of $\boldsymbol{\beta}^*$, to be defined. Intuitively, in view of Banach's fixed point theorem, the sequence of minimizers obtained through iteration of this procedure converges to the global conquer estimator, itself converging to $\boldsymbol{\beta}^*$. In the limit of increasing iterations, there is no information loss over the oracle procedure with access to all data simultaneously, in spite of the data being distributed. The theoretical results of this section establish the delicate choices of h, b, and the number of iterations in order for the *synthesis error* to match the statistical error of the global conquer estimator.

Under the conditional quantile model (1), the generic data vector $(y, \boldsymbol{x})$ can be written in a linear form $y = \boldsymbol{x}^{\mathrm{T}} \boldsymbol{\beta}^* + \varepsilon$, where the model error ε satisfies $Q_\tau(\varepsilon | \boldsymbol{x}) = 0$. Let $f_{\varepsilon|\boldsymbol{x}}(\cdot)$ be the conditional density function of ε given $\boldsymbol{x}$. Given i.i.d. observations $\{(y_i, \boldsymbol{x}_i)\}_{i=1}^N$, we write $\varepsilon_i = y_i - \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}^*$, satisfying $\mathbb{P}(\varepsilon_i \le 0 | \boldsymbol{x}_i) = \tau$.

To investigate the statistical properties of $\widetilde{\boldsymbol{\beta}}^{(1)}$, we impose some regularity conditions.

(C1). There exist $\bar{f} \ge \underline{f} > 0$ such that $\underline{f} \le f_{\varepsilon|\boldsymbol{x}}(0) \le \bar{f}$ almost surely (over $\boldsymbol{x}$). Moreover, there exists some $l_0 > 0$ such that $|f_{\varepsilon|\boldsymbol{x}}(u) - f_{\varepsilon|\boldsymbol{x}}(v)| \le l_0 |u - v|$ for all $u, v \in \mathbb{R}$ almost surely.
(C2). $K(\cdot)$ is a symmetric and non-negative kernel that satisfies $\kappa_2 := \int_{-\infty}^{\infty} u^2 K(u) \, \mathrm{d}u < \infty$, $\kappa_u := \sup_{u \in \mathbb{R}} K(u) < \infty$ and $\kappa_l := \min_{|u| \le 1} K(u) > 0$.
(C3). The predictor $\boldsymbol{x} \in \mathbb{R}^p$ is *sub-Gaussian*: there exists $\upsilon_1 > 0$ such that $\mathbb{P}(|\boldsymbol{z}^{\mathrm{T}} \boldsymbol{u}| \ge \upsilon_1 t) \le 2e^{-t^2/2}$ for every unit vector $\boldsymbol{u} \in \mathbb{S}^{p-1}$ and $t \ge 0$, where $\boldsymbol{z} = \Sigma^{-1/2} \boldsymbol{x}$ and $\Sigma = \mathbb{E}(\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}})$ is positive definite.

Condition (C1) imposes regularity conditions on the conditional density function. These are standard in quantile regression. In (C2), the requirement $\min_{|u| \le 1} K(u) > 0$ is for technical simplicity and can be relaxed to $\min_{|u| \le c} K(u) > 0$ for some $c \in (0, 1)$, which will only change the constant terms in all of our theoretical results. In particular, for kernels that are compactly supported on $[-1, 1]$, we may choose $c = 1/2$ and assume $\kappa_l = \min_{|u| \le 1/2} K(u) > 0$ instead. Distributions with heavier tails than Gaussian on $\boldsymbol{x}$ are excluded by Condition (C3) in order to guarantee exponential-type concentration bounds for estimators of quantile regression coefficients.

For some radii $r, r_* > 0$, define the events

$$\mathcal{E}_0(r) = \big\{\widetilde{\boldsymbol{\beta}}^{(0)} \in \Theta(r)\big\} \quad \text{and} \quad \mathcal{E}_*(r_*) = \big\{\|\nabla \widehat{\mathcal{Q}}_h(\boldsymbol{\beta}^*)\|_\Omega \le r_*\big\}, \tag{10}$$

where

$$\Theta(r) := \{\boldsymbol{\beta} \in \mathbb{R}^p : \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_\Sigma \le r\} \quad \text{and} \quad \Omega := \Sigma^{-1}.$$

In particular, $\mathcal{E}_0(r)$ is a "good" event on which the initial estimator $\widetilde{\boldsymbol{\beta}}^{(0)}$ falls into a local neighborhood $\Theta(r)$ around $\boldsymbol{\beta}^*$. Recall that $\mathbf{H} = \mathbb{E}\{f_{\varepsilon|\boldsymbol{x}}(0)\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\}$ is the Hessian of the population quantile loss $\boldsymbol{\beta} \mapsto \mathbb{E}\{\rho_\tau(y - \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta})\}$ at $\boldsymbol{\beta}^*$. The following theorem provides the statistical properties of $\widetilde{\boldsymbol{\beta}}^{(1)}$.

Theorem 1 *Assume Conditions (C1)–(C3) hold, and let $\mathcal{E}_0(r_0)$ and $\mathcal{E}_*(r_*)$ be the events defined in (10) for some $r_0 \gtrsim r_* > 0$. Let $x > 0$, and suppose the bandwidths $b \ge h > 0$ satisfy $\max\{r_0, \sqrt{(p+x)/n}\} \lesssim b \lesssim 1$ and $\sqrt{(p+x)/N} \lesssim h$. Conditioned on the event $\mathcal{E}_0(r_0) \cap \mathcal{E}_*(r_*)$, the one-step estimator $\widetilde{\boldsymbol{\beta}}^{(1)}$ defined in (9) satisfies*

$$\|\widetilde{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}^*\|_\Sigma \lesssim \left(\sqrt{\frac{p+x}{nb}} + \sqrt{\frac{p+x}{Nh}} + b\right) \cdot r_0 + r_* \tag{11}$$

and

$$\|\,\mathbf{H}(\widetilde{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}^*) + \nabla \widehat{\mathcal{Q}}_h(\boldsymbol{\beta}^*)\,\|_\Omega \lesssim \left(\sqrt{\frac{p+x}{nb}} + \sqrt{\frac{p+x}{Nh}} + b\right) \cdot r_0 \tag{12}$$

with probability at least $1 - 3e^{-x}$.

Equation (11) is the prediction error for the estimator obtained from running a single iteration of our proposed method, while equation (12) provides bounds on a linear Bahadur representation of the estimator, used later for detailed statistical inference on $\boldsymbol{\beta}^*$ or functionals thereof.

Before proceeding, we first discuss some implications of Theorem 1. The parameter r_0 captures the convergence rate of the initial estimator $\widetilde{\boldsymbol{\beta}}^{(0)}$. It can be constructed either on a single local machine that has access to n observations or via averaging all the local estimators. The former is communication-free, while the latter usually improves the statistical accuracy at the cost of one round of communication. Therefore, we may expect a conservative convergence rate of the initial estimator, which is of order $\sqrt{p/n}$. In this case, the rate $r_0 \asymp \sqrt{p/n}$ is sub-optimal compared to that of the global QR estimator $\widehat{\boldsymbol{\beta}}$ in (3) or the conquer estimator $\widehat{\boldsymbol{\beta}}^{\mathrm{cq}}$ in (5). Large sample properties of $\widehat{\boldsymbol{\beta}}^{\mathrm{cq}}$ have been examined by Fernandes, Guerre and Horta (2021) when p is fixed, and by He *et al.* (2021) under the increasing-p regime. According to the latter, the expected prediction error of $\widehat{\boldsymbol{\beta}}^{\mathrm{cq}}$, namely $\|\widehat{\boldsymbol{\beta}}^{\mathrm{cq}} - \boldsymbol{\beta}^*\|_\Sigma$, is primarily determined by $\|\nabla \widehat{\mathcal{Q}}_h(\boldsymbol{\beta}^*)\|_\Omega$ which is of order $\sqrt{p/N} + h^2$; see Lemma 16 in the Appendix. Therefore, the second term on the right-hand side of (11) corresponds to the optimal statistical rate, provided that $(p/N)^{1/2} \lesssim h \lesssim (p/N)^{1/4}$ when all the data are used. Turning to the first term, we see that with properly chosen bandwidths b and h, say $b \asymp (p/n)^{1/3}$ and $h \asymp (p/N)^{1/3}$, the one-step estimator $\widetilde{\boldsymbol{\beta}}^{(1)}$ refines the statistical accuracy of $\widetilde{\boldsymbol{\beta}}^{(0)}$ by a factor of order $(p/n)^{1/3}$.

We can repeat the one-step procedure in (9) using $\widetilde{\boldsymbol{\beta}}^{(1)}$ as an initial estimator, thereby obtaining $\widetilde{\boldsymbol{\beta}}^{(2)}$. After T iterations, we denote the resulting distributed QR estimator by $\widetilde{\boldsymbol{\beta}}^{(T)}$. Since the statistical error is reduced by a factor of $(p/n)^{1/3}$, with high probability, at each iteration, we expect that after $\Omega\big(\lceil \log(m)/\log(n/p) \rceil\big)$ iterations, the communication-efficient distributed estimator $\widetilde{\boldsymbol{\beta}}^{(T)}$ will achieve the same convergence rate as the global estimator $\widehat{\boldsymbol{\beta}}$ or $\widehat{\boldsymbol{\beta}}^{\mathrm{cq}}$.

We formally describe the above iterative procedure as follows, starting at iteration 0 with an initial estimate $\widetilde{\boldsymbol{\beta}}^{(0)}$. At iteration $t = 1, 2, \ldots$, construct the shifted conquer loss function

$$\widetilde{\mathcal{Q}}^{(t)}(\boldsymbol{\beta}) = \widehat{\mathcal{Q}}_{1,b}(\boldsymbol{\beta}) - \big\langle \nabla \widehat{\mathcal{Q}}_{1,b}(\widetilde{\boldsymbol{\beta}}^{(t-1)}) - \nabla \widehat{\mathcal{Q}}_h(\widetilde{\boldsymbol{\beta}}^{(t-1)}), \boldsymbol{\beta} \big\rangle, \tag{13}$$

yielding $\widetilde{\boldsymbol{\beta}}^{(t)}$ that minimizes $\widetilde{\mathcal{Q}}^{(t)}(\cdot)$, that is,

$$\widetilde{\boldsymbol{\beta}}^{(t)} \in \operatorname*{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^p} \widetilde{\mathcal{Q}}^{(t)}(\boldsymbol{\beta}). \tag{14}$$

As before, $b \geq h > 0$ are the local and global bandwidths, respectively. The details are described in Algorithm 1. Notably, the shifted loss $\widetilde{\mathcal{Q}}^{(t)}(\cdot)$ $(t \geq 1)$ is twice-differentiable, convex and (provably) locally strongly convex. To solve the shifted conquer loss minimization problem in (14), in Section A.1 of the Appendix, we describe a gradient descent (GD) algorithm modified by the application of a Barzilai-Borwein step (Barzilai and Borwein, 1988). Such a first-order algorithm is computationally scalable to large dimensions.

In reminiscence of the classical one-step estimator of Bickel (1975), we may instead seek an approximate solution to the minimization problem (14) at each iteration by performing one step of Newton's method. At iteration t, $\nabla \widetilde{\mathcal{Q}}^{(t)}(\widetilde{\boldsymbol{\beta}}^{(t-1)}) = \nabla \widehat{\mathcal{Q}}_h(\widetilde{\boldsymbol{\beta}}^{(t-1)})$ and $\nabla^2 \widetilde{\mathcal{Q}}^{(t)}(\widetilde{\boldsymbol{\beta}}^{(t-1)}) = \nabla^2 \widehat{\mathcal{Q}}_{1,b}(\widetilde{\boldsymbol{\beta}}^{(t-1)})$. Thus, starting with an initialization $\overline{\boldsymbol{\beta}}^{(0)}$, the Newton step computes the update $\overline{\boldsymbol{\beta}}^{(t)} = \overline{\boldsymbol{\beta}}^{(t-1)} - \{\nabla^2 \widehat{\mathcal{Q}}_{1,b}(\overline{\boldsymbol{\beta}}^{(t-1)})\}^{-1} \nabla \widehat{\mathcal{Q}}_h(\overline{\boldsymbol{\beta}}^{(t-1)})$ for $t = 1, 2, \ldots$. At each iteration, the above one-step update essentially performs a Newton-type step based on $\overline{\boldsymbol{\beta}}^{(t-1)}$. While computationally advantageous, the desirable statistical properties of this one-step estimator rely on uniform convergence of the sample Hessian, which typically requires stronger scaling with the sample size.

Theorem 2 below provides the statistical properties of the distributed conquer estimator $\widetilde{\boldsymbol{\beta}}^{(T)}$, including high probability bounds on both estimation error and Bahadur linearization error. The latter serves as an intermediate step for establishing the asymptotic distribution of $\widetilde{\boldsymbol{\beta}}^{(T)}$. Similar results can be obtained for $\overline{\boldsymbol{\beta}}^{(T)}$. In fact, the analysis in this case is much simpler due to the closed-form expression, and is therefore omitted.

Theorem 2 *Assume the same set of conditions in Theorem 1. Then, conditioned on* $\mathcal{E}_0(r_0) \cap \mathcal{E}_*(r_*)$*, the distributed estimator* $\widetilde{\boldsymbol{\beta}}^{(T)}$ *with* $T \gtrsim \log(r_0/r_*)/\log(1/b)$ *satisfies*

$$\begin{aligned} \|\widetilde{\boldsymbol{\beta}}^{(T)} - \boldsymbol{\beta}^*\|_\Sigma &\lesssim r_*, \\ \| \mathbf{H}(\widetilde{\boldsymbol{\beta}}^{(T)} - \boldsymbol{\beta}^*) + \nabla \widehat{\mathcal{Q}}_h(\boldsymbol{\beta}^*) \|_\Omega &\lesssim \big\{ \sqrt{(p+x)/(nb)} + \sqrt{(p+x)/(Nh)} + b \big\} \cdot r_*, \end{aligned} \tag{15}$$

with probability at least $1 - (2T+1)e^{-x}$.

Algorithm 1 Distributed Quantile Regression via Convolution Smoothing.

Input: data batches $\{(y_i, \boldsymbol{x}_i)\}_{i\in\mathcal{I}_j}$, $j = 1, \ldots, m$, stored on m local machines, quantile level $\tau \in (0,1)$, bandwidths $b, h > 0$, initialization $\widetilde{\boldsymbol{\beta}}^{(0)}$, maximum number of iterations T, $g_0 = 1$.

1: **for** $t = 1, 2 \ldots, T$ **do**
2: Broadcast $\widetilde{\boldsymbol{\beta}}^{(t-1)}$ to all local machines.
3: **for** $j = 1, \ldots, m$ **do**
4: Compute $\nabla\widehat{\mathcal{Q}}_{j,h}(\widetilde{\boldsymbol{\beta}}^{(t-1)})$ on the jth local machine, and send it to the master (first) machine.
5: **end for**
6: Compute the global gradient $\nabla\widehat{\mathcal{Q}}_h(\widetilde{\boldsymbol{\beta}}^{(t-1)}) = (1/m)\sum_{j=1}^m \nabla\widehat{\mathcal{Q}}_{j,h}(\widetilde{\boldsymbol{\beta}}^{(t-1)})$ and its ℓ_∞-norm $g_t = \|\nabla\widehat{\mathcal{Q}}_h(\widetilde{\boldsymbol{\beta}}^{(t-1)})\|_\infty$ on the master.
7: **if** $g_t > g_{t-1}$ or $g_t < 10^{-5}$ **break**
8: **otherwise** Compute $\nabla\widehat{\mathcal{Q}}_{1,b}(\widetilde{\boldsymbol{\beta}}^{(t-1)})$, and solve $\widetilde{\boldsymbol{\beta}}^{(t)} \in \operatorname{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} \widetilde{\mathcal{Q}}^{(t)}(\boldsymbol{\beta})$ on the master.
9: **end for**

Output: $\widetilde{\boldsymbol{\beta}}^{(T)}$.

Remark 3 *From the proof of Theorem 2, we see that the multi-round estimate $\widetilde{\boldsymbol{\beta}}^{(t)}$ after t iterations satisfies with high probability that*

$$\|\widetilde{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^*\|_2 \lesssim \delta^t \cdot r_0 + r_* \quad \textit{with} \quad \delta = \sqrt{\frac{p}{nb}} + \sqrt{\frac{p}{Nh}} + b,$$

where r_0 and r_* represent, respectively, the initial convergence rate and the global rate (attainable by the centralized estimator). This result also characterizes the trade-off between communication cost and estimation accuracy. After running the algorithm for t rounds, the communication cost for each local machine/node is $\mathcal{O}(pt)$. On the other hand, since the statistical limit of distributed estimation is determined by r_*, we need as many as $\mathcal{O}(\log(r_*/r_0)/\log(1/\delta))$ communication rounds for the proposed distributed estimator to achieve the optimal rate, resulting in a total communication cost $\mathcal{O}(p\log(r_*/r_0)/\log(1/\delta))$ for each local machine. Ignoring logarithmic factors, the above parameters (r_0, r_*, δ) will be taken as

$$r_0 \asymp \sqrt{\frac{p}{n}}, \quad r_* \asymp \sqrt{\frac{p}{N}}, \quad \textit{and} \quad \delta \asymp \left(\frac{p}{n}\right)^{1/3}.$$

Now let us discuss the construction of the initial estimator $\widetilde{\boldsymbol{\beta}}^{(0)}$. Using a local sample from a single source, we can take $\widetilde{\boldsymbol{\beta}}^{(0)}$ to be either the standard QR estimator (Koenker and Bassett, 1978) or the conquer estimator described in Section 2.1. That is,

$$\widetilde{\boldsymbol{\beta}}^{(0)} \in \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{1}{n}\sum_{i\in\mathcal{I}_1} \rho_\tau(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}) \quad \text{or} \quad \widetilde{\boldsymbol{\beta}}^{(0)} \in \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} \widehat{\mathcal{Q}}_{1,b}(\boldsymbol{\beta}), \tag{16}$$

where $b > 0$ is the local bandwidth. In the diverging-p regime, explicit high probability error bounds for the QR and conquer estimators can be found in Pan and Zhou (2021) and He *et al.* (2021), respectively.

Theorem 4 *Assume that Conditions (C1)–(C3) hold, and choose the bandwidths $b, h > 0$ as $b \asymp \{(p + \log(n \log m))/n\}^{1/3}$ and $h \asymp \{(p + \log(n \log m))/N\}^{\gamma}$ for any $\gamma \in [1/3, 1/2]$. Moreover, suppose the sample size per source satisfies $n \gtrsim p + \log(n \log m)$. Then, starting at iteration 0 with an initial estimate $\widetilde{\boldsymbol{\beta}}^{(0)}$ given in (16), the multi-round distributed estimator $\widetilde{\boldsymbol{\beta}} = \widetilde{\boldsymbol{\beta}}^{(T)}$ with $T \asymp \lceil \frac{\log(m)}{\log(n/p)} \rceil$ satisfies*

$$\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_{\Sigma} \lesssim \sqrt{\frac{p + \log(n \log m)}{N}} \tag{17}$$

and

$$\left\| \mathbf{H}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) + \frac{1}{N} \sum_{i=1}^{N} \{\overline{K}(-\varepsilon_i/h) - \tau\} \boldsymbol{x}_i \right\|_{\Omega} \lesssim \frac{(p + \log(n \log m))^{5/6}}{n^{1/3} N^{1/2}} + \frac{p + \log(n \log m)}{N h^{1/2}} \tag{18}$$

with probability at least $1 - Cn^{-1}$, where $m = N/n$ is the number of sources.

Remark 5 *Guided by the theoretically "optimal" choice of the local and global bandwidths stated in Theorem 4, in practice we suggest to choose*

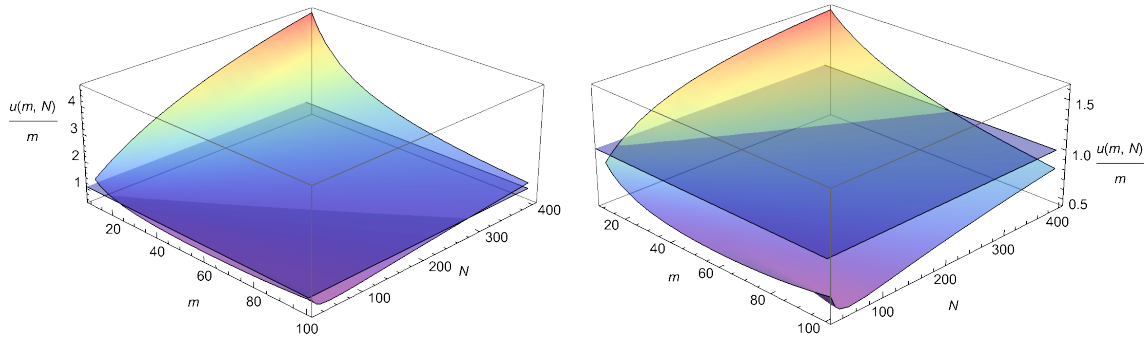
$$b = c \cdot \left(\frac{p + \log n}{n} \right)^{1/3} \quad \text{and} \quad h = c \cdot \left(\frac{p + \log N}{N} \right)^{1/3}, \tag{19}$$

for some positive constant c. For preprocessed data that has constant-level scales, we may choose c from $\{0.5, 1, 2.5, 5\}$ using a validation set. More generally, we consider a heuristic, dynamic method for choosing c. To solve the optimization problem in (14), Section A.1 describes a quasi-Newton-type algorithm, namely the gradient descent with step size automatically determined by the Barzilai-Borwein method. At each iteration, we set c in (19) to be the minimum between the sample standard deviation and the median absolute deviation (multiplied by 1.4826) of the residuals from the previous iterate. The resulting estimate is then scale-invariant.

The scaling condition $n \gtrsim p + \log(n \log m)$, while not as easy to parse as the $m \lesssim \sqrt{N}$ condition implicated by simple meta analyses, is appreciably less stringent. To visualize this constraint, we introduce the function

$$u(m, N) := \frac{N}{p + \log(N/m) + \log(\log m)}$$

so that the knife-edge permissible value of m arises when $m = u(m, N)$. This fixed point equation is thus solved when $u(m, N)/m = 1$. Figure 1 plots $u(m, N)/m$ against m and N for $p = n/10$ and $p = n/2$. The permissible scaling of m with N is the curve traced out by the intersection of $u(m, N)/m$ with the constant function, taking value 1 for all values of the argument. From Figure 1, the permissible scaling of m with N is visibly faster than $\sqrt{N}$. By comparing Figures 1(a) and (b), we see that this scaling is made more severe by proportional increases in p.



(a) (b)

Figure 1: Plot of $u(m, N)/m$ against m and N overlaid with the constant function to indicate the fixed point of $u(m, N)$ for (a) $p = n/10$ and (b) $p = n/2$.

Using a local estimator as the initialization is most efficient in terms of storage, communication, and computational complexity. Alternatively, one can use the so called divide-and-conquer (meta analysis) estimator based on simply averaging the local QR estimators (Volgushev, Chao and Cheng, 2019) as the initialization. This improves the statistical stability at the cost of one round of communication. For $j = 1, \ldots, m$, define the local empirical loss functions $\widehat{\mathcal{Q}}_j(\boldsymbol{\beta}) = (1/n) \sum_{i \in \mathcal{I}_j} \rho_\tau(y_i - \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta})$, and the corresponding local QR estimators $\widehat{\boldsymbol{\beta}}_j^{\mathrm{loc}} \in \operatorname{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^p} \widehat{\mathcal{Q}}_j(\boldsymbol{\beta})$. Estimators obtained from the separate sources are combined after one round of communication to construct a global estimator, namely the divide-and-conquer quantile regression (DC-QR) estimator

$$\widehat{\boldsymbol{\beta}}^{\mathrm{dc}} = \frac{1}{m} \sum_{j=1}^{m} \widehat{\boldsymbol{\beta}}_j^{\mathrm{loc}}. \tag{20}$$

For quantile regression, Volgushev, Chao and Cheng (2019) derived the estimation error of $\widehat{\boldsymbol{\beta}}^{\mathrm{dc}}$ when the (random) covariates have fixed dimension p and are uniformly bounded, that is, $\max_{1 \le i \le n} \|\boldsymbol{x}_i\|_2 \le c_p$ for some $c_p > 0$. Under regularity conditions that are similar to Condition (C1), Theorem 3.1 therein implies

$$\|\widehat{\boldsymbol{\beta}}^{\mathrm{dc}} - \widehat{\boldsymbol{\beta}}\|_2 = \mathcal{O}_{\mathbb{P}}\left(\frac{\log N}{n} + \frac{(\log N)^{7/4}}{n^{1/4} N^{1/2}} \right) + o_{\mathbb{P}}(N^{-1/2})$$

as long as $m = o(N/\log N)$. If communication constraints allow, we recommend using the DC-QR estimator $\widehat{\boldsymbol{\beta}}^{\mathrm{dc}}$ as the initial estimator, and setting $T = \max\{\lceil \log m \rceil, 2\}$ in Algorithm 1. The whole procedure hence requires at most $T + 1$ communication rounds.

In Section 4.2, we demonstrate via numerical studies that the bias of the DC-QR estimator is visibly larger than the bias of the proposed distributed conquer estimator under extreme quantile regression models with heteroscedastic errors. As a result, confidence sets based on a normal approximation to the DC-QR Wald statistic are susceptible to severe undercoverage in linear heteroscedastic models.

Remark 6 *Just as the statistical aspects of extreme value theory are challenged by the limitation of data beyond extreme thresholds, QR coefficients at extreme quantiles are notoriously hard to estimate. Section D of the Appendix details a minor adaptation of our procedure which improves its performance at extreme quantile levels.*

2.3 Distributed inference

2.3.1 Wald-type confidence sets

With a view to more detailed statistical inference beyond point estimation, we first establish a distributional approximation in the form of a Berry-Esseen bound. This forms the basis for a Wald test, which can be inverted to give confidence sets for $\boldsymbol{\beta}^*$ and linear functionals thereof. Construction of the pivotal test statistic relies on a consistent estimator of the asymptotic variance, which is typically obtained using a nonparametric estimate of the conditional density function of the response given the covariates.

Theorem 7 *Under the same set of conditions in Theorem 4, the distributed conquer estimator $\widetilde{\boldsymbol{\beta}} = \widetilde{\boldsymbol{\beta}}^{(T)}$ satisfies*

$$\sup_{x \in \mathbb{R},\, \boldsymbol{a} \in \mathbb{R}^p} \big| \mathbb{P}\big\{ N^{1/2} \boldsymbol{a}^{\mathrm{T}} (\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) / \sigma_{\tau,h} \le x \big\} - \Phi(x) \big| \\ \lesssim \frac{p + \log(n \log m)}{(Nh)^{1/2}} + N^{1/2} h^2 + \frac{(p + \log(n \log m))^{5/6}}{n^{1/3}}, \tag{21}$$

where $\sigma_{\tau,h}^2 = \boldsymbol{a}^{\mathrm{T}} \mathbf{H}^{-1} \mathbb{E}[\{\overline{K}(-\varepsilon/h) - \tau\}^2 \boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}] \mathbf{H}^{-1} \boldsymbol{a}$ and $\Phi(\cdot)$ is the standard normal distribution function. In particular, under the scaling $p + \log(\log m) = o(\min\{n^{2/5}, N^{3/8}\})$, the distributed estimator $\widetilde{\boldsymbol{\beta}}$ with bandwidths $b \asymp \{(p + \log(n \log m))/n\}^{1/3}$ and $h \asymp \{(p + \log(n \log m))/N\}^{2/5}$ satisfies

$$N^{1/2} \sigma_{\tau,h}^{-1} \boldsymbol{a}^{\mathrm{T}} (\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \xrightarrow{\mathrm{d}} \mathcal{N}(0,1) \quad \text{and} \quad \frac{N^{1/2} \boldsymbol{a}^{\mathrm{T}} (\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)}{(\boldsymbol{a}^{\mathrm{T}} \mathbf{H}^{-1} \Sigma \mathbf{H}^{-1} \boldsymbol{a})^{1/2}} \xrightarrow{\mathrm{d}} \mathcal{N}\big(0, \tau(1-\tau)\big)$$

uniformly over $\boldsymbol{a} \in \mathbb{R}^p$ as $n \to \infty$, where $\xrightarrow{\mathrm{d}}$ is a shorthand for convergence in distribution.

The accuracy of the normal approximation hinges on both the global and local bandwidths, and on the scaling of m with N and p. The role of b is via (15), in view of which, the upper bound in Theorem 7 is of order

$$(p+x)^{1/2} \left(\sqrt{\frac{p+x}{nb}} + b + \sqrt{\frac{p+x}{Nh}} \right) + N^{1/2} h^2,$$

where $x = \log(n \log m)$. Minimizing as a function of (h, b) delivers the rate in Theorem 7 by taking

$$b \asymp \left(\frac{p+x}{n} \right)^{1/3} \quad \text{and} \quad h \asymp \left(\frac{p+x}{N} \right)^{2/5}.$$

To our knowledge, Theorem 7 is the first Berry-Esseen inequality with explicit error bounds depending on both n and p in a distributed setting.

We first describe methods that use the normal distribution with estimated variance for calibration. Let $\widetilde{\boldsymbol{\beta}} = \widetilde{\boldsymbol{\beta}}^{(T)}$ be the communication-efficient estimator discussed in the previous subsection. Under mild conditions, Theorem 7 establishes the asymptotic normality that for every $1 \le j \le p$,

$$\frac{N^{1/2}(\widetilde{\beta}_j - \beta_j^*)}{(\mathbf{H}^{-1}\Sigma\mathbf{H}^{-1})_{jj}^{1/2}} \xrightarrow{\mathrm{d}} \mathcal{N}\big(0, \tau(1-\tau)\big),$$

where $\mathbf{H} = \mathbb{E}\{f_{\varepsilon|\boldsymbol{x}}(0)\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\}$ and $\Sigma = \mathbb{E}(\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}})$. The problem is then reduced to estimating the pointwise variance $(\mathbf{H}^{-1}\Sigma\mathbf{H}^{-1})_{jj}$, i.e., the j^{th} diagonal entry of $\mathbf{H}^{-1}\Sigma\mathbf{H}^{-1}$.

To this end, define the residual function and fitted residuals as

$$\varepsilon_i(\boldsymbol{\beta}) = y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} \ \text{ for } \ \boldsymbol{\beta} \in \mathbb{R}^p \ \text{ and } \ \widehat{\varepsilon}_i = \varepsilon_i(\widetilde{\boldsymbol{\beta}}) = y_i - \boldsymbol{x}_i^{\mathrm{T}}\widetilde{\boldsymbol{\beta}}, \ \ i = 1, \ldots, N. \tag{22}$$

In a nondistributed setting, the $p \times p$ Hessian matrix $\mathbf{H}$ can be estimated by the following variant of Powell's kernel-type estimator (Powell, 1991)

$$\widehat{\mathbf{H}}_b = \frac{1}{m}\sum_{j=1}^{m}\widehat{\mathbf{H}}_{j,b} \ \text{ with } \ \widehat{\mathbf{H}}_{j,b} = \frac{1}{nb}\sum_{i\in\mathcal{I}_j}\phi\big(\widehat{\varepsilon}_i/b\big)\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}, \ j = 1, \ldots, m, \tag{23}$$

where $\phi(\cdot)$ is the standard normal density function and $b > 0$ is a bandwidth that may differ from the previous one. Moreover, define $\widehat{\Sigma} = (1/m)\sum_{j=1}^{m}\widehat{\Sigma}_j$ and $\widehat{\Sigma}_b(\tau) = (1/m)\sum_{j=1}^{m}\widehat{\Sigma}_{j,b}(\tau)$, where

$$\widehat{\Sigma}_j = \frac{1}{n}\sum_{i\in\mathcal{I}_j}\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}} \quad \text{and} \quad \widehat{\Sigma}_{j,b}(\tau) = \frac{1}{n}\sum_{i\in\mathcal{I}_j}\big\{\overline{K}(-\widehat{\varepsilon}_i/b) - \tau\big\}^2\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}. \tag{24}$$

Computing the full matrix estimators $\widehat{\mathbf{H}}_b$ and $\widehat{\Sigma}$ or $\widehat{\Sigma}_b(\tau)$ requires each machine to communicate $p \times p$ local estimators $\widehat{\mathbf{H}}_{j,b}$ and $\widehat{\Sigma}_j$ or $\widehat{\Sigma}_{j,b}(\tau)$ to the master machine. This incurs excessive communication cost.

To achieve a trade-off between communication efficiency and statistical accuracy, we instead use a local pointwise variance estimator

$$\tau(1-\tau)\big(\widehat{\mathbf{H}}_{1,b}^{-1}\widehat{\Sigma}_1\widehat{\mathbf{H}}_{1,b}^{-1}\big)_{jj} \quad \text{or} \quad \big(\widehat{\mathbf{H}}_{1,b}^{-1}\widehat{\Sigma}_{1,b}(\tau)\widehat{\mathbf{H}}_{1,b}^{-1}\big)_{jj}. \tag{25}$$

For the latter, note that $\widehat{\Sigma}_{1,b}(\tau)$ can be viewed as a sample analog of $\mathbb{E}\{\overline{K}(-\varepsilon/b) - \tau\}^2\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}$, which is closely related to the asymptotic variance of $\widetilde{\boldsymbol{\beta}}$ as revealed by Theorem 7. Moreover, as discussed in Fernandes, Guerre and Horta (2021), the width of a confidence interval based on $\widehat{\mathbf{H}}_{1,b}^{-1}\widehat{\Sigma}_{1,b}(\tau)\widehat{\mathbf{H}}_{1,b}^{-1}$ for any element of $\boldsymbol{\beta}^*$ is asymptotically narrower than that based on the naïve variance estimator $\tau(1-\tau)\,\widehat{\mathbf{H}}_{1,b}^{-1}\widehat{\Sigma}_1\widehat{\mathbf{H}}_{1,b}^{-1}$.

For every $\boldsymbol{\beta} \in \mathbb{R}^p$, define the matrix-valued function

$$\widehat{\mathbf{H}}_{1,b}(\boldsymbol{\beta}) = \frac{1}{nb}\sum_{i\in\mathcal{I}_1}\phi\big(\varepsilon_i(\boldsymbol{\beta})/b\big)\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}},$$

where $\varepsilon_i(\boldsymbol{\beta}) = y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}$ are as in (22). Under this notation, $\widehat{\mathbf{H}}_{1,b} = \widehat{\mathbf{H}}_{1,b}(\widetilde{\boldsymbol{\beta}})$. The next result provides a uniform convergence result for $\widehat{\mathbf{H}}_{1,b}(\boldsymbol{\beta})$ over $\boldsymbol{\beta}$ in a local neighborhood of $\boldsymbol{\beta}^*$. For any symmetric matrix $\mathbf{A} \in \mathbb{R}^{p\times p}$, we use $\|\cdot\|_\Omega$ ($\Omega = \Sigma^{-1}$) to denote the relative operator norm, that is, $\|\mathbf{A}\|_\Omega = \|\Sigma^{-1/2}\mathbf{A}\Sigma^{-1/2}\|_2$. With this notation, we have $\|\Sigma\|_\Omega = 1$.

Proposition 8 *Conditions (C1)–(C3) ensure that, for any $r, x > 0$,*

$$\sup_{\boldsymbol{\beta}\in\Theta(r)} \|\widehat{\mathbf{H}}_{1,b}(\boldsymbol{\beta}) - \mathbf{H}\|_\Omega \lesssim \sqrt{\frac{p\log n + x}{nb}} + b + r \tag{26}$$

with probability at least $1-3e^{-x}$ as long as $n \gtrsim p+x$ and $b \gtrsim (p\log n+x)/n$. In addition, if $f'_{\varepsilon|\boldsymbol{x}}(\cdot)$ is Lipschitz continuous, that is, $|f'_{\varepsilon|\boldsymbol{x}}(u) - f'_{\varepsilon|\boldsymbol{x}}(v)| \le l_1|u-v|$ for all $u, v \in \mathbb{R}$ almost surely (over $\boldsymbol{x}$), then $\sup_{\boldsymbol{\beta}\in\Theta(r)} \|\widehat{\mathbf{H}}_{1,b}(\boldsymbol{\beta}) - \mathbf{H}\|_\Omega \lesssim \sqrt{(p\log n + x)/(nb)} + b^2 + r$ with high probability.

Theorem 9 *Under the same set of assumptions in Theorem 4, the local estimators $\widehat{\Sigma}_1$ and $\widehat{\mathbf{H}}_{1,b}$ with $b \asymp \{p\log(n)/n\}^{1/3}$ satisfy the bounds*

$$\|\widehat{\Sigma}_1 - \Sigma\|_\Omega \lesssim (p+\log n)^{1/2} n^{-1/2} \quad \textit{and} \quad \|\widehat{\mathbf{H}}_{1,b} - \mathbf{H}\|_\Omega \lesssim (p\log n)^{1/3} n^{-1/3}$$

with probability at least $1 - Cn^{-1}$ as long as $n \gtrsim p\log n$.

In a simpler case where $f_{\varepsilon|\boldsymbol{x}}(0) = f_\varepsilon(0)$ is independent of $\boldsymbol{x}$, we have $\mathbf{H}^{-1}\Sigma\mathbf{H}^{-1} = \{f_\varepsilon(0)\}^{-2}\Omega$ so that it suffices to estimate the univariate density function $f_{\varepsilon|\boldsymbol{x}}(\cdot)$ at 0. Arguably the most commonly used method is the following kernel density estimator:

$$\widehat{f}_\varepsilon(0) = \frac{1}{Nb}\sum_{i=1}^{N} K\big(\widehat{\varepsilon}_i/b\big) = \frac{1}{m}\sum_{j=1}^{m} \widehat{f}_{\varepsilon,j}(0), \tag{27}$$

where $\widehat{f}_{\varepsilon,j}(0) = (nb)^{-1}\sum_{i\in\mathcal{I}_j} K(\widehat{\varepsilon}_i/b)$ for $j = 1,\ldots,m$. Therefore, $\widehat{f}_\varepsilon(0)$ can be easily computed in a distributed manner. For convenience, we use the standard normal density as kernel function and the rule-of-thumb bandwidth by Hall and Sheather (1988), that is,

$$b_N^{\mathrm{rot}} = N^{-1/3}\cdot\Phi^{-1}(1-\alpha/2)^{2/3}\left\{\frac{1.5\cdot\phi(\Phi^{-1}(\tau))^2}{2\Phi^{-1}(\tau)^2+1}\right\}^{1/3},$$

where α is a prepecified probability of miscoverage. For the kernel matrix estimators $\widehat{\mathbf{H}}_{1,b}$ and $\widehat{\Sigma}_{1,b}(\tau)$ defined in (23) and (24), we use the same local bandwidth b as in Algorithm 1 for efficient distributed quantile regression. The corresponding normal-based confidence intervals for β_j^* $(j = 1,\ldots,p)$ are given by

$$\Big[\widetilde{\beta}_j - \Phi^{-1}(1-\alpha/2)\cdot\widehat{\sigma}_j\cdot N^{-1/2},\ \ \widetilde{\beta}_j + \Phi^{-1}(1-\alpha/2)\cdot\widehat{\sigma}_j\cdot N^{-1/2}\Big], \tag{28}$$

where $\widehat{\sigma}_j = (\widehat{\mathbf{H}}_{1,b}^{-1}\widehat{\Sigma}_{1,b}(\tau)\widehat{\mathbf{H}}_{1,b}^{-1})_{jj}^{1/2}$, $\sqrt{\tau(1-\tau)}\,(\widehat{\mathbf{H}}_{1,b}^{-1}\widehat{\Sigma}_1\widehat{\mathbf{H}}_{1,b}^{-1})_{jj}^{1/2}$ or $\widehat{f}_\varepsilon(0)^{-1}(\widehat{\Sigma}_1)_{jj}^{1/2}\sqrt{\tau(1-\tau)}$. The first two variance estimates are preferred under general heteroscedastic models in which $\mathbf{H} = \mathbb{E}\{f_{\varepsilon|\boldsymbol{x}}(0)\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\}$ no longer takes the form $f_\varepsilon(0)\cdot\Sigma$.

Remark 10 *The construction of normal-based confidence intervals as in (28) depends crucially on the asymptotic variance estimation. The validity of $\widehat{\sigma}_j = \widehat{f}_\varepsilon(0)^{-1}(\widehat{\Sigma}_1)_{jj}^{1/2}\sqrt{\tau(1-\tau)}$ relies on the assumption that $\mathbf{H} = \mathbb{E}\{f_{\varepsilon|\boldsymbol{x}}(0)\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\}$ takes the form $f_\varepsilon(0)\cdot\Sigma$. This holds*

trivially when the model error ε and covariates $\boldsymbol{x}$ are independent, which is arguably too restrictive in the context of quantile regression. More generally, let us consider a standard location-scale model $y = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^* + \sigma(\boldsymbol{x}) \cdot e$, where $e \sim f_e(\cdot)$ is independent of $\boldsymbol{x}$ and $\sigma(\cdot)$ is a non-negative function. In this case, we have $\varepsilon = \sigma(\boldsymbol{x}) \cdot e$, whose conditional and unconditional densities at 0 are $f_{\varepsilon|\boldsymbol{x}}(0) = f_e(0)/\sigma(\boldsymbol{x})$ and $f_\varepsilon(0) = f_e(0) \cdot \mathbb{E}\{1/\sigma(\boldsymbol{x})\}$. This reveals that $\mathbf{H} = f_e(0) \cdot \mathbb{E}\{\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}/\sigma(\boldsymbol{x})\}$ and $f_\varepsilon(0) \cdot \Sigma$ are generally unequal, and therefore the use of $\widehat{\sigma}_j = (\widehat{\mathbf{H}}_{1,b}^{-1}\widehat{\Sigma}_{1,b}(\tau)\widehat{\mathbf{H}}_{1,b}^{-1})_{jj}^{1/2}$ or $\widehat{\sigma}_j = \sqrt{\tau(1-\tau)}\,(\widehat{\mathbf{H}}_{1,b}^{-1}\widehat{\Sigma}_1\widehat{\mathbf{H}}_{1,b}^{-1})_{jj}^{1/2}$. is more robust and preferable under heteroscedastic models.

2.3.2 Score-type confidence sets

While the Wald test inverts to give explicit confidence intervals as in equation (28), confidence sets based on other types of test acknowledge that the set of parameter values consistent with the data need not form an interval.

For some $k = 1, \ldots, p$, consider the hypothesis

$$H_0^k : \beta_k^* = c_k \quad \text{versus} \quad H_1^k : \beta_k^* \neq c_k, \tag{29}$$

where c_k is a predetermined constant. Let $\widetilde{\boldsymbol{\beta}}_{H^k} = (\widetilde{\beta}_{H^k,1}, \ldots, \widetilde{\beta}_{H^k,p})^{\mathrm{T}} \in \mathbb{R}^p$ denote the distributed quantile regression estimator with its kth coordinate constrained at the hypothesized value, i.e., $\widetilde{\beta}_{H^k,k} = c_k$. To construct a score test, define the gradient

$$\widehat{\boldsymbol{S}} = (\widehat{S}_1, \ldots, \widehat{S}_p)^{\mathrm{T}} = N \cdot \nabla \widehat{\mathcal{Q}}_h(\widetilde{\boldsymbol{\beta}}_{H^k}) = \sum_{j=1}^{m} \sum_{i \in \mathcal{I}_j} \widehat{\xi}_i \boldsymbol{x}_i, \tag{30}$$

where $\widehat{\xi}_i = \overline{K}\{(\boldsymbol{x}_i^{\mathrm{T}}\widetilde{\boldsymbol{\beta}}_{H^k} - y_i)/h\} - \tau$. Under the null hypothesis H_0^k, it is reasonable to expect the t-statistic $\widehat{T}_k$, which is defined as $N^{1/2}\widehat{S}_k$ divided by the estimated standard deviation, to be asymptotically normally distributed. We can write the t-statistic in terms of the self-normalized sum $\widehat{T}_k$ as (Efron, 1969):

$$\widehat{T}_k = \frac{\widehat{S}_k/\widehat{V}_k}{\sqrt{\{N - (\widehat{S}_k/\widehat{V}_k)^2\}/(N-1)}}, \tag{31}$$

where $\widehat{S}_k = \sum_{j=1}^m \sum_{i \in \mathcal{I}_j} \widehat{\xi}_i x_{ik}$ and $\widehat{V}_k^2 = \sum_{j=1}^m \sum_{i \in \mathcal{I}_j} (\widehat{\xi}_i x_{ik})^2$. This representation has the advantage that the quantities $\widehat{S}_k$ and $\widehat{V}_k$ can be calculated in a distributed manner without information loss.

Write $\xi_i = \overline{K}(-\varepsilon_i/h) - \tau$ and $\mu_k = \mathbb{E}(\xi_i x_{ik})$, where $\varepsilon_i = y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*$. Denote the "oracle" version of $\widehat{T}_k$ by T_k:

$$T_k = \frac{N^{-1/2}\sum_{i=1}^N (\xi_i x_{ik} - \mu_k)}{\sqrt{(N-1)^{-1}\sum_{i=1}^N (\xi_i x_{ik} - N^{-1}\sum_{\ell=1}^N \xi_\ell x_{\ell k})^2}} = \frac{S_k/V_k}{\sqrt{\{N - (S_k/V_k)^2\}/(N-1)}},$$

where $S_k = \sum_{i=1}^N (\xi_i x_{ik} - \mu_k)$ and $V_k^2 = \sum_{i=1}^N (\xi_i x_{ik} - \mu_k)^2$. Note that S_k is a sum of independent zero-mean random variables. Asymptotic properties of the self-normalized

sum S_k/V_k have been well established in the literature (de la Paña, Lai and Shao, 2009). On writing $\widehat{T}_k$ more explicitly as $\widehat{T}_k(c_k)$, we define the α-level confidence set associated with the score test as

$$\big\{\, c_k : \Phi^{-1}(\alpha/2) \le \widehat{T}_k(c_k) \le \Phi^{-1}(1-\alpha/2) \,\big\}. \tag{32}$$

This will often, but need not always, deliver intervals. The possibility of non-interval confidence sets should be viewed as an advantage, as exemplified by Fieller's problem (Fieller, 1954). The disadvantage of using the score statistic for constructing confidence sets is that $\widehat{T}_k(c_k)$ has to be evaluated for a multitude of c_k values, in practice over a fine grid of points. The computational burden of this is considerable relative to the Wald construction in Section 2.3.1.

2.3.3 Resampling-based confidence sets

An alternative widely used approach treats an interval as the primary mode of inference rather than the significance test and constructs the former directly by resampling methods such as the bootstrap. Resampling approaches typically provide tighter confidence limits than the Wald-based interval due to their implicit higher-order accuracy over limiting distributional approximations. However, the computational burden is high in the present context.

Recall from Theorem 4 that the multi-round distributed estimator $\widetilde{\boldsymbol{\beta}} = (\widetilde{\beta}_1, \ldots, \widetilde{\beta}_p)^{\mathrm{T}}$ admits the following asymptotic linear (Bahadur) representation:

$$N^{1/2}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) = -\mathbf{H}^{-1} \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \big\{ \overline{K}(-\varepsilon_i/h) - \tau \big\} \boldsymbol{x}_i + o_{\mathbb{P}}(1).$$

Motivated by this asymptotic representation, Belloni et al. (2017) suggested and proved the validity of the *multiplier score bootstrap*, which is based on randomly perturbing the asymptotic linear forms of the nonlinear quantile regression estimators. Intuitively, the distribution of $N^{1/2}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)$ can be approximately estimated by the bootstrap draw of

$$N^{1/2}(\widetilde{\boldsymbol{\beta}}^\flat - \widetilde{\boldsymbol{\beta}}) := -\widehat{\mathbf{H}}^{-1} \frac{1}{\sqrt{N}} \sum_{i=1}^{N} e_i \big\{ \overline{K}(-\widehat{\varepsilon}_i/h) - \tau \big\} \boldsymbol{x}_i, \tag{33}$$

where $e_1, \ldots, e_N$ are i.i.d. standard normal random variables, $\widehat{\mathbf{H}}$ denotes a generic (consistent) estimator of $\mathbf{H}$, and $\widehat{\varepsilon}_i$ are fitted residuals. In the distributed framework, each bootstrap draw requires one round of communication. The composite communication cost can be exorbitant when the number of bootstrap replications is large, say 1000 or 2000.

Recently, Yu, Chao and Cheng (2020) proposed two bootstrap methods for constructing simultaneous confidence intervals with distributed data. To operationalize their proposals in the present context, define $\widehat{\boldsymbol{\xi}}_i = \{\overline{K}(-\widehat{\varepsilon}_i/h) - \tau\}\boldsymbol{x}_i$ for $i = 1, \ldots, N$, and let $\widehat{\mathbf{H}}_1$ be a local estimator of $\mathbf{H}$ using the n samples on the first machine. For example, $\widehat{\mathbf{H}}_1$ can be taken as either $\widehat{\mathbf{H}}_{1,b}$ given in (23) or $\widehat{f}_\varepsilon(0) \cdot \widehat{\Sigma}_1$. Then, consider the following two multiplier bootstrap statistics

$$\boldsymbol{w}^\sharp = (w_1^\sharp, \ldots, w_p^\sharp)^{\mathrm{T}} = -\widehat{\mathbf{H}}_1^{-1} \frac{1}{\sqrt{m}} \sum_{j=1}^{m} e_j \cdot n^{1/2} \nabla \widehat{\mathcal{Q}}_{j,h}(\widetilde{\boldsymbol{\beta}}) \tag{34}$$

and

$$\boldsymbol{w}^\flat = (w_1^\flat, \ldots, w_p^\flat)^{\mathrm{T}} = -\widehat{\mathbf{H}}_1^{-1} \frac{1}{\sqrt{n+m-1}} \left\{ \sum_{i=1}^{n} e_i \cdot \widehat{\boldsymbol{\xi}}_i + \sum_{j=2}^{m} e_{n+j-1} \cdot n^{1/2} \nabla \widehat{\mathcal{Q}}_{j,h}(\widetilde{\boldsymbol{\beta}}) \right\}, \quad (35)$$

both of which only require one additional round of communication, and therefore are communication-efficient. As before, $e_1, \ldots, e_{n+m-1}$ are i.i.d. standard normal variables.

For any $q \in (0,1)$ and $1 \leq j \leq p$, let $\mathrm{c}_j^\sharp(q)$ and $\mathrm{c}_j^\flat(q)$ be the (conditional) q-quantiles of $w_j^\sharp$ and $w_j^\flat$, respectively, defined as $\mathrm{c}_j^\sharp(q) = \inf\{t \in \mathbb{R} : \mathbb{P}^*(w_j^\sharp \leq t) \geq q\}$ and $\mathrm{c}^\flat(q) = \inf\{t \in \mathbb{R} : \mathbb{P}^*(w_j^\flat \leq t) \geq q\}$, where $\mathbb{P}^*(\cdot) = \mathbb{P}(\cdot \,|\, y_1, \boldsymbol{x}_1, \ldots, y_N, \boldsymbol{x}_N)$ denotes the conditional probability given the observed samples. The ensuing bootstrap confidence intervals for β_j^* $(j = 1, \ldots, p)$ are given by

$$\left[\widetilde{\beta}_j - \frac{\mathrm{c}_j^\sharp(1-\alpha/2)}{\sqrt{N}}, \widetilde{\beta}_j - \frac{\mathrm{c}_j^\sharp(\alpha/2)}{\sqrt{N}} \right] \text{ and } \left[\widetilde{\beta}_j - \frac{\mathrm{c}_j^\flat(1-\alpha/2)}{\sqrt{N}}, \widetilde{\beta}_j - \frac{\mathrm{c}_j^\flat(\alpha/2)}{\sqrt{N}} \right]. \quad (36)$$

Our simulations in Section 4.2 show that the two bootstrap methods have nearly identical performance when m is large, while the latter is more stable and thus preferable when m is relatively small. We leave the theoretical analysis of these distributed bootstrap methods in the future as a significant amount of additional work is still needed.

2.4 Comparison with prior work

The problem of distributed quantile regression has been considered in two earlier papers. Volgushev, Chao and Cheng (2019) established the statistical properties of the estimator obtained by averaging m local estimators, each constructed according to equation (3). The single round of communication and direct use of the check function means that there are no tuning parameters. However, as indicated in Section 1, the permissible scaling of m with N required to ensure the optimal statistical properties is restrictive, and violation of this constraint leads to under-coverage of resulting confidence sets. Under-coverage is particularly severe under the highly plausible scenario in which the quantile regression error depends on the covariates. See Section 4 for an empirical demonstration.

For M-estimation with a convex loss, Chen, Liu and Zhang (2021) proposed a general multi-round distributed procedure paired with stochastic gradient descent. When applied to quantile regression, their approach is a variant of stochastic subgradient descent. For minimizing a convex but non-differentiable function, subgradient methods typically exhibit very slow (sublinear) convergence and hence are not computationally stable. This explains the unpopularity of subgradient approaches among other computational methods for quantile regression. Theoretically, their distributed QR estimator needs a sufficiently large local sample size—namely $n \gtrsim (Np)^{1/2} \log(N)$, to achieve the optimal rate $\mathcal{O}_{\mathbb{P}}(\sqrt{p/N})$; see Theorem 4.7 therein. In addition to the suboptimal scaling, Chen, Liu and Zhang (2021) only derived the convergence rate for point estimation without the uncertainty quantification sought in the present work.

The same authors (Chen, Liu and Zhang, 2019) proposed a procedure specifically for distributed quantile regression. Their smoothed loss function is closer to that of Horowitz (1998) and incompatible with the ideas of Jordan, Lee and Yang (2019) and Wang *et*

al. (2017) due to violation of the uniform Lipschitz continuity condition of the second derivative. They instead exploit a representation of the estimator in terms of estimator-dependent "sufficient statistics". Since the representation is not of closed form, an iterative approach is required, using the estimate at iteration t to update the sufficient statistics at each component source. While the limited communication improves the permissible scaling of m with N over the approach of Volgushev, Chao and Cheng (2019), the construction is such that m $p \times p$ matrices (and other quantities), are communicated at each iteration. Communication of hessian matrices is generally viewed as too communication intensive, particularly when p is large.

We note that none of these approaches is generalizable to the sparse high-dimensional setting. The penalization required to enforce sparsity in high dimensions exacerbates bias so that meta-analysis hinges of the ability to de-bias such estimators prior to aggregation. Attempts to construct de-biased estimators for quantile regression have, so far, relied on an unrealistic assumption that the quantile regression error is independent of the covariates. The key representation used by Chen, Liu and Zhang (2019) is violated upon penalization of their smooth quantile regression estimator, and suffers from singularity when $p > n$ if penalization is not applied.

An anonymous reviewer pointed out concurrent work by Jiang and Yu (2021) (the first version of our manuscript dates back to late September, 2020) who also proposed communication-efficient algorithms for distributed quantile regression by means of convolution smoothing. The main difference between this work and ours concerns the theoretical aspects. Under similar regularity and moment conditions, we provide explicit non-asymptotic concentration bounds as well as Berry-Esseen-type bounds for normal approximation. These results complement the conventional $\mathcal{O}_{\mathbb{P}}$ statements in Jiang and Yu (2021). To achieve the global convergence rate in low-dimensions, Theorem 3.2 in Jiang and Yu (2021) requires (p, n, N) to satisfy $n = N^r$ and $p \asymp N^c$ for some $0 < r \leq 1$ and $0 < c < \min(3/8, r)$, while our result (Theorem 4) only requires $n \gtrsim p + \log\log(N/n)$. In high-dimensions, from Theorem 4.1 in Jiang and Yu (2021) and its proof we see that the dimension p cannot exceed the sample size N in the sense that $p \asymp N^c$ for some $c \in (0, 1)$. Our results, detailed in Section 3, show that the penalized distributed QR estimator achieves the global rate under the sample size requirements $n \gtrsim s^2 \log p$ and $N \gtrsim s^3 \log p$, which considerably relax those in Jiang and Yu (2021).

3. Distributed Penalized Quantile Regression in High Dimensions

In this section, we consider quantile regression in high-dimensional sparse models with distributed data. In such models, the total number of predictors p can be very large, while the number of important predictors is significantly smaller. As before, assume that the data set $\{(y_i, \boldsymbol{x}_i)\}_{i=1}^N$ with $N = n \cdot m$ is distributed across m sources, so that each source j contributes n i.i.d. observations $\mathcal{D}_j = \{(y_i, x_i)\}_{i \in \mathcal{I}_j}$ indexed by $\mathcal{I}_j$. Assume further that the sparsity $\|\boldsymbol{\beta}^*\|_0 := \sum_{j=1}^p \mathbb{1}(\beta_j^* \neq 0)$ is at most s, which is much smaller than the local sample size, that is, $s = o(n)$.

3.1 Penalized conquer with distributed data

To fit sparse models in high dimensions, the use of ℓ_1 penalization has become a common practice since the seminal work of Tibshirani (1996). The ℓ_1-penalized quantile regression (ℓ_1-QR) estimator is defined as

$$\widehat{\boldsymbol{\beta}} \in \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{1}{N}\sum_{i=1}^{N} \rho_\tau(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}) + \lambda\cdot\|\boldsymbol{\beta}\|_1 = \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} \widehat{\mathcal{Q}}(\boldsymbol{\beta}) + \lambda\cdot\|\boldsymbol{\beta}\|_1, \tag{37}$$

where $\lambda > 0$ is a regularization parameter. Statistical properties and computational methods for ℓ_1-QR have been well studied in the past decade; see, for example, Wang, Li and Jiang (2007), Wu and Lange (2008), Li and Zhu (2008), Belloni and Chernozhukov (2011), Wang, Wu and Li (2012), Yi and Huang (2017) and Gu *et al.* (2018). Recently, Tan, Wang and Zhou (2022) studied the ℓ_1-penalized conquer (ℓ_1-conquer) estimator, which is a solution to the following optimization problem

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \underbrace{\frac{1}{N}\sum_{i=1}^{N} (\rho_\tau * K_h)(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})}_{\widehat{\mathcal{Q}}_h(\boldsymbol{\beta})} + \lambda\cdot\|\boldsymbol{\beta}\|_1, \tag{38}$$

where $K(\cdot)$ is a non-negative kernel and $h > 0$ is the bandwidth.

Notably, the smoothed loss function $\widehat{\mathcal{Q}}_h(\cdot)$ is (provably) strongly convex in a local neighborhood of $\boldsymbol{\beta}^*$ with high probability. With a proper initialization, the corresponding optimization problem with ℓ_1-penalization can be efficiently solved via first-order algorithms. In a distributed setting, we extend the iterative algorithm in Section 2 as follows. Let $\widetilde{\boldsymbol{\beta}}^{(0)} \in \mathbb{R}^p$ be an initial regularized estimator. Denote by $\widetilde{\mathcal{Q}}(\boldsymbol{\beta}) = \widehat{\mathcal{Q}}_{1,b}(\boldsymbol{\beta}) - \langle \nabla\widehat{\mathcal{Q}}_{1,b}(\widetilde{\boldsymbol{\beta}}^{(0)}) - \nabla\widehat{\mathcal{Q}}_h(\widetilde{\boldsymbol{\beta}}^{(0)}), \boldsymbol{\beta}\rangle$ the same shifted conquer loss as in (8), where b and h are the local and global bandwidths. Analogously to (9), the communication-efficient penalized conquer estimator is defined as

$$\widetilde{\boldsymbol{\beta}}^{(1)} \in \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} \widetilde{\mathcal{Q}}(\boldsymbol{\beta}) + \lambda\cdot\|\boldsymbol{\beta}\|_1, \tag{39}$$

where $\lambda > 0$ is a regularization parameter. Optimization problem (39) is convex, which we solve using a local adaptive majorize-minimize algorithm detailed in Section A.2 of the Appendix.

Let $\mathcal{S} \subseteq \{1,\ldots,p\}$ be the support of $\boldsymbol{\beta}^*$, and assume that data are generated from a sparse conditional quantile model (1) with $|\mathcal{S}| \le s$. Define the ℓ_1-cone

$$\Lambda = \Lambda(s,p) = \big\{\boldsymbol{\beta}\in\mathbb{R}^p : \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_1 \le 4s^{1/2}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_\Sigma\big\}. \tag{40}$$

Given $r > 0$ and $\lambda_* > 0$, we define the "good" events

$$\mathcal{E}_0(r) = \big\{\widetilde{\boldsymbol{\beta}}^{(0)} \in \Theta(r)\cap\Lambda\big\} \quad\text{and}\quad \mathcal{E}_*(\lambda_*) = \big\{\|\nabla\widehat{\mathcal{Q}}_h(\boldsymbol{\beta}^*) - \nabla\mathcal{Q}_h(\boldsymbol{\beta}^*)\|_\infty \le \lambda_*\big\}, \tag{41}$$

which, with slight abuse of notation, extend those given in (10) to the high-dimensional setting.

In the following, we first establish upper bounds for the ℓ_1- and ℓ_2-errors of the one-step penalized estimator $\widetilde{\boldsymbol{\beta}}^{(1)}$, provided that the initial estimator $\widetilde{\boldsymbol{\beta}}^{(0)}$ falls in a local neighborhood

of $\boldsymbol{\beta}^*$. In parallel to Condition (C3), we impose the following moment condition on the high-dimensional random vector $\boldsymbol{x} \in \mathbb{R}^p$ of covariates.

(C4). The predictor $\boldsymbol{x} = (x_1, \ldots, x_p)^{\mathrm{T}} \in \mathbb{R}^p$ (with $x_1 \equiv 1$) has bounded components and uniformly bounded kurtosis. That is, there exists $B \geq 1$ such that $\max_{1 \leq j \leq p} |x_j| \leq B$ almost surely, and $\mu_4 := \sup_{\boldsymbol{u} \in \mathbb{S}^{p-1}} \mathbb{E}(\boldsymbol{z}^{\mathrm{T}} \boldsymbol{u})^4 < \infty$, where $\boldsymbol{z} = \Sigma^{-1/2} \boldsymbol{x}$ and $\Sigma = (\sigma_{jk})_{1 \leq j,k \leq p} = \mathbb{E}(\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}})$ is positive definite. Write $\sigma_u = \max_{1 \leq j \leq p} \sigma_{jj}^{1/2}$ and $\lambda_l = \lambda_{\min}(\Sigma) \in (0, 1]$. For convenience, we assume $\lambda_l = 1$.

For technical reasons, the bounded covariates assumption is also imposed in Wang *et al.* (2017) and Jordan, Lee and Yang (2019) for sparse linear regression and generalized linear models.

Theorem 11 *Assume Conditions (C1), (C2), and (C4) hold. For $\delta \in (0, 1)$ and $r_0, \lambda_* > 0$, let $b \geq h > 0$ and $\lambda = 2.5(\lambda_* + \varrho) > 0$ satisfy*

$$\varrho \asymp \max\left[\left\{\frac{1}{b}\sqrt{\frac{\log(p/\delta)}{n}} + \frac{1}{h}\sqrt{\frac{\log(p/\delta)}{N}}\right\} s^{1/2} r_0,\ s^{-1/2}\big(b r_0 + h^2\big)\right] \quad \text{and} \quad s^{1/2}\lambda \lesssim b \lesssim 1.$$

Conditioned on the event $\mathcal{E}_0(r_0) \cap \mathcal{E}_*(\lambda_*)$, the one-step estimator $\widetilde{\boldsymbol{\beta}}^{(1)}$ defined in (39) *satisfies $\widetilde{\boldsymbol{\beta}}^{(1)} \in \Lambda$ and*

$$\|\widetilde{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}^*\|_\Sigma \lesssim \left\{\frac{s}{b}\sqrt{\frac{\log(p/\delta)}{n}} + b + \frac{s}{h}\sqrt{\frac{\log(p/\delta)}{N}}\right\} r_0 + s^{1/2}\lambda_* + h^2 \tag{42}$$

with probability at least $1 - \delta$.

In Theorem 11, the prespecified parameter $r_0 > 0$ quantifies the accuracy of the initial regularized QR estimator $\widetilde{\boldsymbol{\beta}}^{(0)}$ under ℓ_2-norm. Using a subsample of size n to construct such an estimator, the nearly minimax-optimal rate is $\sqrt{s\log(p)/n}$ (Belloni and Chernozhukov, 2011; Wang and He, 2021). With a suitable choice for the regularization weight λ, we ensure that $\widetilde{\boldsymbol{\beta}}^{(1)}$ must lie in the restricted set Λ, and bandwidths $b, h > 0$, the estimation error of $\widetilde{\boldsymbol{\beta}}^{(1)}$ is of the order

$$\underbrace{\left\{\frac{s}{b}\sqrt{\frac{\log(p/\delta)}{n}} + b + \frac{s}{h}\sqrt{\frac{\log(p/\delta)}{N}}\right\}}_{\text{contraction factor}} r_0 + \underbrace{s^{1/2}\lambda_* + h^2}_{\text{near-optimal rate}}.$$

As we shall see, the second term is related to the near-optimal rate when the entire dataset is used. The first term involves a contraction factor that is of the order $\frac{s}{b}\sqrt{\log(p/\delta)/n} + b + \frac{s}{h}\sqrt{\log(p/\delta)/N}$. With sufficiently many samples per source—namely, $n \gtrsim s^2 \log(p/\delta)$, the above one-step estimation procedure, which uses one round of communication, improves the statistical accuracy of $\widetilde{\boldsymbol{\beta}}^{(0)}$ as long as $s\sqrt{\log(p/\delta)/n} \lesssim b \lesssim 1$ and $s\sqrt{\log(p/\delta)/N} \lesssim h \lesssim 1$.

Next, we describe an iterative, multi-round procedure for estimating a sparse $\boldsymbol{\beta}^* \in \mathbb{R}^p$ in a distributed setting. Let $\widehat{Q}_{j,b}(\cdot)$ and $\widehat{Q}_{j,h}(\cdot)$, $j = 1, \ldots, m$, be the local empirical loss functions given in (7). At iteration 0, the first (master) machine computes an initial estimator $\widetilde{\boldsymbol{\beta}}^{(0)}$ as well as $\nabla \widehat{\mathcal{Q}}_{1,b}(\widetilde{\boldsymbol{\beta}}^0)$, and broadcast $\widetilde{\boldsymbol{\beta}}^{(0)}$ to all local machines. For $j = 1, \ldots, m$, the

jth local machine then computes gradients $\nabla \widehat{\mathcal{Q}}_{j,h}(\widetilde{\boldsymbol{\beta}}^{(0)})$, which are then transmitted back to the first. At iteration $t = 1, 2, \ldots, T$, the first machine solves the ℓ_1-penalized shifted conquer loss minimization

$$\widetilde{\boldsymbol{\beta}}^{(t)} \in \operatorname*{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^p} \underbrace{\widehat{\mathcal{Q}}_{1,b}(\boldsymbol{\beta}) - \big\langle \nabla \widehat{\mathcal{Q}}_{1,b}(\widetilde{\boldsymbol{\beta}}^{(t-1)}) - \nabla \widehat{\mathcal{Q}}_{h}(\widetilde{\boldsymbol{\beta}}^{(t-1)}), \boldsymbol{\beta} \big\rangle}_{=: \widetilde{\mathcal{Q}}^{(t)}(\boldsymbol{\beta})} + \lambda_t \cdot \|\boldsymbol{\beta}\|_1, \tag{43}$$

where $\nabla \widehat{\mathcal{Q}}_{h}(\widetilde{\boldsymbol{\beta}}^{(t-1)}) = (1/m) \sum_{j=1}^m \nabla \widehat{\mathcal{Q}}_{j,h}(\widetilde{\boldsymbol{\beta}}^{(t-1)})$, and $\lambda_t > 0$ are regularization parameters.

Theorem 12 *Assume Conditions (C1), (C2), and (C4) hold. Given $\delta \in (0,1)$, choose the local and global bandwidths as*

$$b \asymp s^{1/2} \big\{ \log(p/\delta)/n \big\}^{1/4} \quad \text{and} \quad h \asymp \big\{ s \log(p/\delta)/N \big\}^{1/4}. \tag{44}$$

For $r_0, \lambda_* > 0$, write $r_* = s^{1/2}\lambda_*$ and set $\lambda_t = 2.5(\lambda_* + \varrho_t) > 0$ ($t \geq 1$) with

$$\varrho_t \asymp \max \Big\{ \gamma^t s^{-1/2} r_0 + \gamma s^{-1/2}(r_* + h^2) \mathbb{1}(t \geq 2),\ \sqrt{\log(p/\delta)/N} \Big\},$$

where $\gamma = \gamma(s, p, n, N, \delta) \asymp s^{1/2} \max\{\log(p/\delta)/n, s \log(p/\delta)/N\}^{1/4}$. Let the sample size per source and total sample size satisfy $n \gtrsim s^2 \log(p/\delta)$ and $N \gtrsim s^3 \log(p/\delta)$, so that $\gamma < 1$. Moreover, assume $r_0 \lesssim \min\{1, (m/s)^{1/4}\}$ and $r_* \lesssim b$. Then, conditioned on the event $\mathcal{E}_*(\lambda_*) \cap \mathcal{E}_0(r_0)$, the T^{th} iterate $\widetilde{\boldsymbol{\beta}}^{(T)}$ with $T \gtrsim \log(r_0/r_*)/\log(1/\gamma)$ satisfies

$$\|\widetilde{\boldsymbol{\beta}}^{(T)} - \boldsymbol{\beta}^*\|_\Sigma \lesssim s^{1/2}\lambda_* + h^2 \quad \text{and} \quad \|\widetilde{\boldsymbol{\beta}}^{(T)} - \boldsymbol{\beta}^*\|_1 \lesssim s\lambda_* + s^{1/2}h^2 \tag{45}$$

with probability at least $1 - T\delta$.

According to Theorem 12, the success of the iterative procedure described above relies on a sufficiently accurate initial estimator $\widetilde{\boldsymbol{\beta}}^{(0)}$. For example, we may choose $\widetilde{\boldsymbol{\beta}}^{(0)}$ to be a local ℓ_1-conquer estimator

$$\widetilde{\boldsymbol{\beta}}^{(0)} \in \operatorname*{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^p} \widehat{\mathcal{Q}}_{1,b_0}(\boldsymbol{\beta}) + \lambda_0 \cdot \|\boldsymbol{\beta}\|_1, \tag{46}$$

or a local ℓ_1-QR estimator which is a minimizer of the program

$$\operatorname*{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^p} \frac{1}{n} \sum_{i \in \mathcal{I}_1} \rho_\tau(y_i - \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}) + \lambda_0 \cdot \|\boldsymbol{\beta}\|_1. \tag{47}$$

High probability estimation error bounds for ℓ_1-QR were derived by Belloni and Chernozhukov (2011), Wang (2013), and more recently by Wang and He (2021) under weaker assumptions. The estimation error for ℓ_1-conquer is provided by the following result, which is a variant of Theorem 4.1 in Tan, Wang and Zhou (2022).

Proposition 13 *Assume Conditions (C1), (C2) and (C4) hold. For $\delta \in (0,1)$, set the regularization parameter $\lambda_0 \asymp \sqrt{\tau(1-\tau)\log(p/\delta)/n}$. Provided that $\sqrt{s \log(p/\delta)/n} \lesssim b_0 \lesssim 1$, the local ℓ_1-conquer estimator $\widetilde{\boldsymbol{\beta}}^{(0)}$ given in (46) satisfies*

$$\|\widetilde{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^*\|_\Sigma \lesssim s^{1/2}\lambda_0 + b_0^2 \tag{48}$$

with probability at least $1 - \delta$. If in addition $b_0 \lesssim \{s \log(p/\delta)/n\}^{1/4}$, then $\widetilde{\boldsymbol{\beta}}^{(0)} \in \Lambda$.

With the above preparations, we are now ready to state the estimator error bound for the distributed regularized conquer estimator $\widetilde{\boldsymbol{\beta}}^{(T)}$ in high dimensions.

Theorem 14 *Assume Conditions (C1), (C2) and (C4) hold, and that the data are generated from a sparse conditional quantile model* (1) *with* $\|\boldsymbol{\beta}^*\|_0 \leq s$. *Suppose the sample size per source and total sample size satisfy* $n \gtrsim s^2 \log(p)$ *and* $N \gtrsim s^3 \log(p)$. *Choose the bandwidths* $b, h > 0$ *and regularization parameters* λ_t *(*$t \geq 1$*) as* $b \asymp s^{1/2}\{\log(p)/n\}^{1/4}$, $h \asymp \{s \log(p)/N\}^{1/4}$ *and*

$$\lambda_t \asymp \sqrt{\frac{\log(p)}{N}} + \max\left\{\frac{s^2 \log(p)}{n}, \frac{s^3 \log(p)}{N}\right\}^{t/4} \sqrt{\frac{\log(p)}{n}}.$$

Starting at iteration 0 with an initial estimate $\widetilde{\boldsymbol{\beta}}^{(0)}$ *as described in Proposition 13, the distributed estimator* $\widetilde{\boldsymbol{\beta}} = \widetilde{\boldsymbol{\beta}}^{(T)}$ *with* $T \asymp \lceil \log(m) \rceil$ *communication rounds satisfies the error bounds*

$$\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \lesssim \sqrt{\frac{s \log(p)}{N}} \quad \textit{and} \quad \|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \lesssim s\sqrt{\frac{\log(p)}{N}} \tag{49}$$

with probability at least $1 - C \log(m)/N$.

Theorems 11–14 are non-trivial extensions of Theorem 3 in Wang *et al.* (2017) to the context of quantile regression. The latter can be applied to the squared loss for linear regression and logistic loss for classification. Let $\ell(\cdot)$ be the loss function of interest, and it is assumed therein that

$$|\ell'(u) - \ell'(v)| \leq L|u - v| \;\text{ for any }\; u, v \in \mathbb{R} \quad \text{and} \quad \sup_{u \in \mathbb{R}} |\ell'''(u)| \leq M.$$

The key of the proof is to control the difference between the gradient vectors $\nabla \widetilde{\mathcal{Q}}^{(t)}(\widetilde{\boldsymbol{\beta}}^{(t-1)})$ and $\nabla \widetilde{\mathcal{Q}}^{(t)}(\boldsymbol{\beta}^*)$ at each iteration. For this purpose, the proof of Theorem 3 in Wang *et al.* (2017) is based on the second-order Taylor's series expansion, so that the above parameters L and M arise and are treated as constants. In particular, $M = 0$ for the quadratic loss. In our context, if we take $\ell(\cdot)$ to be the local conquer loss $(\rho_\tau * K_b)(\cdot)$, then it is easy to see that $L \asymp b^{-1}$ and $M \asymp b^{-2}$. Since the bandwidth b decays as a function of (n, p), neither the result nor proof argument in Wang *et al.* (2017) apply to quantile regression even with smoothing. In the Appendix, we provide a self-contained proof of Theorems 11 and 12, which relies on a uniform control of the fluctuations of gradient processes and a restricted strong convexity property for the empirical conquer loss.

3.2 Distributed quantile regression via ADMM

In this subsection, we describe an alternative algorithm based on the alternating direction method of multiplier (ADMM) for penalized quantile regression with distributed data. ADMM, which was first introduced by Douglas and Rachford (1956) and Gabay and Mercier (1976), has a number of successful applications in modern statistical machine learning. We refer to Boyd *et al.* (2011) for a comprehensive review on ADMM. In the context of quantile

regression, Yu, Lin and Wang (2017) and Gu *et al.* (2018) respectively proposed ADMM-based algorithms for fitting penalized QR with both convex and folded-concave penalties.

As argued in Boyd *et al.* (2011), ADMM is well suited for distributed convex optimization problems under minimum structural assumption. For solving penalized QR, in the following we revisit the parallel implementation of the ADMM-based algorithm proposed in Yu, Lin and Wang (2017). Recall that the total dataset $\{(y_i, \boldsymbol{x}_i)\}_{i=1}^N$ with $N = n \cdot m$ is distributed across m sources, each containing a data batch indexed by $\mathcal{I}_j$ $(j = 1, \ldots, m)$. Write

$$\boldsymbol{y} = (y_1, \ldots, y_N)^{\mathrm{T}} = (\boldsymbol{y}_1^{\mathrm{T}}, \ldots, \boldsymbol{y}_m^{\mathrm{T}})^{\mathrm{T}} \quad \text{and} \quad \boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)^{\mathrm{T}} = (\boldsymbol{X}_1^{\mathrm{T}}, \ldots, \boldsymbol{X}_m^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{N \times p},$$

where $\boldsymbol{y}_j = \boldsymbol{y}_{\mathcal{I}_j} \in \mathbb{R}^n$ and $\boldsymbol{X}_j \in \mathbb{R}^{n \times p}$. Under this set of notation, the ℓ_1-QR problem (37) can be recast into an equivalent problem

$$\underset{\boldsymbol{r}_j, \boldsymbol{\beta}_j, \boldsymbol{\beta}}{\text{minimize}} \left\{ \sum_{j=1}^m \rho_\tau(\boldsymbol{r}_j) + \lambda_N \|\boldsymbol{\beta}\|_1 \right\} \text{ such that } \boldsymbol{y}_j - \boldsymbol{X}_j \boldsymbol{\beta}_j = \boldsymbol{r}_j, \quad \boldsymbol{\beta}_j = \boldsymbol{\beta}, \quad j = 1, \ldots, m,$$

where $\lambda_N = N\lambda$. Here we write $\rho_\tau(\boldsymbol{r}) = \sum_{i=1}^n \rho_\tau(r_i)$ for $\boldsymbol{r} = (r_1, \ldots, r_n)^{\mathrm{T}}$. To solve this linearly constrained optimization problem, the ADMM updates at iteration $k = 0, 1, \ldots$ are

$$\begin{aligned}
\boldsymbol{\beta}^{k+1} &= \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \frac{m\gamma}{2} \big\| \boldsymbol{\beta} - \bar{\boldsymbol{\beta}}^k - \bar{\boldsymbol{\delta}}^k / \gamma \big\|_2^2 + \lambda_N \|\boldsymbol{\beta}\|_1 \right\}, && (50) \\
\boldsymbol{r}_j^{k+1} &= \underset{\boldsymbol{r}_j}{\operatorname{argmin}} \left\{ \rho_\tau(\boldsymbol{r}_j) + \frac{\gamma}{2} \big\| \boldsymbol{y}_j - \boldsymbol{X}_j \boldsymbol{\beta}_j^k + \boldsymbol{u}_j^k / \gamma - \boldsymbol{r}_j \big\|_2^2 \right\}, && (51) \\
\boldsymbol{\beta}_j^{k+1} &= (\boldsymbol{X}_j^{\mathrm{T}} \boldsymbol{X}_j + \mathbf{I}_p)^{-1} \big\{ \boldsymbol{X}_j^{\mathrm{T}} \big(\boldsymbol{y}_j - \boldsymbol{r}_j^{k+1} + \boldsymbol{u}_j^k / \gamma \big) - \boldsymbol{\delta}_j^k / \gamma + \boldsymbol{\beta}^{k+1} \big\}, \\
\boldsymbol{u}_j^{k+1} &= \boldsymbol{u}_j^k + \gamma \big(\boldsymbol{y}_j - \boldsymbol{X}_j \boldsymbol{\beta}_j^{k+1} - \boldsymbol{r}_j^{k+1} \big), \\
\boldsymbol{\delta}_j^{k+1} &= \boldsymbol{\delta}_j^k + \gamma \big(\boldsymbol{\beta}_j^{k+1} - \boldsymbol{\beta}^{k+1} \big),
\end{aligned}$$

where $\bar{\boldsymbol{\beta}}^k = (1/m) \sum_{j=1}^m \boldsymbol{\beta}_j^k$, $\bar{\boldsymbol{\delta}}^k = (1/m) \sum_{j=1}^m \boldsymbol{\delta}_j^k$, and $\gamma > 0$ is the augmentation parameter. In particular, the $\boldsymbol{\beta}$-update in (50) and the $\boldsymbol{r}$-update in (51) have explicit expressions, which are

$$\begin{aligned}
\boldsymbol{\beta}^{k+1} &= \big(\bar{\boldsymbol{\beta}}^k + \bar{\boldsymbol{\delta}}^k / \gamma - \lambda_N / (m\gamma) \mathbf{1}_p \big)_+ - \big(- \bar{\boldsymbol{\beta}}^k - \bar{\boldsymbol{\delta}}^k / \gamma - \lambda_N / (m\gamma) \mathbf{1}_p \big)_- \quad \text{and} \\
\boldsymbol{r}^{k+1} &= \big(\boldsymbol{y}_j - \boldsymbol{X}_j \boldsymbol{\beta}_j^k + \boldsymbol{u}_j^k / \gamma - \tau \gamma^{-1} \mathbf{1}_n \big)_+ - \big(- \boldsymbol{y}_j + \boldsymbol{X}_j \boldsymbol{\beta}_j^k - \boldsymbol{u}_j^k / \gamma + (\tau - 1) \gamma^{-1} \mathbf{1}_n \big)_+,
\end{aligned}$$

respectively, where $\mathbf{1}_q := (1, \ldots, 1)^{\mathrm{T}} \in \mathbb{R}^q$ for each integer $q \geq 1$.

The above parallel version of the ADMM to solve (37) involves primal variables $\boldsymbol{\beta} \in \mathbb{R}^p$, $(\boldsymbol{r}_1^{\mathrm{T}}, \ldots, \boldsymbol{r}_m^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^N$ and the dual variable $(\boldsymbol{u}_1^{\mathrm{T}}, \ldots, \boldsymbol{u}_m^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^N$. As a general-purpose algorithm, its convergence can be quite slow when applied to large-scale datasets. For example, under a numerical setting with $p = 100$, $N = 30,000$ and $m \in \{1, 10, 100\}$ considered in Yu, Lin and Wang (2017), it takes more than 100 iterations for the parallel implementation of the ADMM to converge. In a distributed framework, this amounts to (at least) 100 communication rounds in order to achieve the desired level of statistical accuracy. At even larger data scales, our numerical results (see Figure 4 below) show evidence that the proposed multi-round, distributed estimator can perform as well as the global estimator within $T = 10$ communication rounds.

4. Numerical Studies

4.1 Distributed quantile regression

Starting with the low-dimensional setting, we compare the proposed multi-round procedure with the following methods: (i) global QR estimator using all of the available $N = mn$ observations; (ii) the averaging-based estimator based on local QR estimators; (iii) the proposed method with $T \in \{1, 4, 10\}$ communication rounds; and (iv) a non-smooth version of the proposed method, which uses the subgradient of the QR loss as the global gradient, with $T \in \{1, 4, 10\}$ communication rounds. We employ the `R` packages `conquer` and `quantreg` to compute the conquer and standard QR estimators, respectively.

As shown in He *et al.* (2021), the performance of conquer is insensitive to the choice of kernel functions, and thus we use the Gaussian kernel wherever smoothing is required. Our proposed method involves an initial estimator $\widetilde{\boldsymbol{\beta}}^{(0)}$ and two smoothing parameters h and b. There are multiple ways to obtain an adequate initialization. For instance, as suggested by Jordan, Lee and Yang (2019), one can use the simple averaging estimator as the initialization, i.e., the average of local QR estimators across m sources. For simplicity, we take $\widetilde{\boldsymbol{\beta}}^{(0)}$ to be a conquer estimator computed based on n independent data points from one source. For the bandwidths, we set $h = 2.5 \cdot \{(p+\log N)/N\}^{1/3}$ and $b = 2.5 \cdot \{(p+\log n)/n\}^{1/3}$ according to the theoretical analysis in Section 2.2.

To generate the data, we consider two types of heteroscedastic models:

1. Linear heteroscedasticity: $y_i = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^* + (0.2x_{ip} + 1)\{\varepsilon_i - F_{\varepsilon_i}^{-1}(\tau)\}$;
2. Quadratic heteroscedasticity: $y_i = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^* + 0.5\{1 + (0.25x_{ip} - 1)^2\}\{\varepsilon_i - F_{\varepsilon_i}^{-1}(\tau)\}$,

where $\boldsymbol{x}_i$ is generated from a multivariate uniform distribution on the cube $3^{1/2} \cdot [-1, 1]^{p+1}$ with covariance matrix $\boldsymbol{\Sigma} = (0.5^{|j-k|})_{1 \le j,k \le p+1}$, and $\boldsymbol{\beta}^* = \mathbf{1}_p$ is a p-vector of ones. The random noise is generated from a t-distribution with 2 degrees of freedom, denoted by t_2. To evaluate the performance across different methods, we report the estimation error under the ℓ_2-norm, i.e., $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2$. Table 1 presents the results when $n = 300$, $p = 10$, $m \in \{50, 100, 200, 400, 600, 1000\}$, and $\tau = 0.8$, averaged over 100 trials. With the same p and τ, we report the results with a fixed total sample size $N = 150,000$, $m = N/n$, and varying local sample size $n \in \{300, 500, 1000, 1500, 3000, 6000\}$ in Table 2.

The global QR estimator, which always has the smallest error as expected, serves as a benchmark for communication-efficient methods. From Table 1, we see that the proposed multi-round distributed estimator yields the best performance among the communication-efficient estimators, and as the number of communication rounds grows, it becomes almost as good as the global QR estimator even though the one-step estimator ($T = 1$) performs rather poorly. The performance of the averaging-based QR is comparable to that of the proposed method when the number of machines m is smaller than the local sample size n. As suggested by the theoretical analyses, when m is larger than n, the proposed method outperforms the averaging-based QR. To highlight the importance of smoothing for distributed quantile regression, we also implement the multi-round procedure using the subgradient of the QR loss, namely, $\nabla\widehat{\mathcal{Q}}(\boldsymbol{\beta}) = (nm)^{-1}\sum_{j=1}^{m}\sum_{i\in\mathcal{I}_j}\{\mathbb{1}(y_i < \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}) - \tau\}\boldsymbol{x}_i$, instead of $\nabla\widehat{\mathcal{Q}}_h(\cdot)$. Note that the estimation error of this subgradient-based method is barely improvable as the number of machines increases. When the number of total samples N is fixed, from Table 2,

we find that the subgradient-based method only performs well if the local sample size is extremely large, which makes all the methods desirable. This demonstrates the importance of smoothing in the context of distributed learning with non-smooth loss functions.

Next, we perform a sensitivity analysis to assess the effect of the initial estimator on the final solution of the proposed method with $T = 10$. To this end, we conduct additional numerical studies where we consider different initial estimators, computed using different sample sizes $n_{\text{init}} = \{150, 300, 500, 1000, 5000\}$. Specifically, we consider the aforementioned linear and quadratic heteroscedastic models with $n = 300$, $p = 10$, $m = 400$, and $\tau = 0.8$. The average estimation error for the proposed method with $T = \{1, 4, 10\}$ and that of global QR estimator are summarized in Table 3. From Table 3, we see that the estimation error for the proposed method with $T = 1$ decreases as we increase the sample size used to calculate the initial estimator. Moreover, we see that implementing the proposed method with $T = \{4, 10\}$ improves the estimation error significantly, and that the estimation error is no longer sensitive to the initial sample size. The proposed method with $T = 10$ yields an estimator that performs as well as the global QR (implemented using dataset from all sources) even when $n_{\text{init}} = 150$. The results suggest that the proposed method is not sensitive to the sample size used to calculate the initial estimator after some rounds of communication.

Table 1: Estimation error under linear and quadratic heteroscedastic models with t_2 noise, averaged over 100 trials. Results for $\tau = 0.8$, $n = 300$, and $p = 10$, across $m = \{50, 100, 200, 400, 600, 1000\}$ are reported.

Linear Heteroscedastic Model with $\tau = 0.8$, $n = 300$, and $p = 10$						
Methods	$m = 50$	$m = 100$	$m = 200$	$m = 400$	$m = 600$	$m = 1000$
`averaging-based QR`	0.077	0.060	0.047	0.041	0.037	0.035
`distributed QR` ($T = 1$)	0.197	0.216	0.223	0.213	0.192	0.198
`distributed QR` ($T = 4$)	0.173	0.223	0.256	0.202	0.187	0.175
`distributed QR` ($T = 10$)	0.259	0.341	0.427	0.313	0.271	0.305
`distributed smoothed QR` ($T = 1$)	0.159	0.163	0.163	0.151	0.138	0.143
`distributed smoothed QR` ($T = 4$)	0.076	0.066	0.051	0.032	0.027	0.029
`distributed smoothed QR` ($T = 10$)	0.075	0.071	0.039	0.027	0.021	0.020
`global QR`	0.069	0.050	0.035	0.025	0.019	0.016
Quadratic Heteroscedastic Model with $\tau = 0.8$, $n = 300$, and $p = 10$						
`averaging-based QR`	0.079	0.063	0.050	0.043	0.038	0.036
`distributed QR` ($T = 1$)	0.206	0.210	0.233	0.204	0.198	0.205
`distributed QR` ($T = 4$)	0.198	0.232	0.281	0.211	0.235	0.184
`distributed QR` ($T = 10$)	0.263	0.401	0.423	0.330	0.396	0.332
`distributed smoothed QR` ($T = 1$)	0.162	0.160	0.163	0.151	0.148	0.148
`distributed smoothed QR` ($T = 4$)	0.079	0.062	0.051	0.033	0.034	0.033
`distributed smoothed QR` ($T = 10$)	0.077	0.057	0.041	0.027	0.024	0.021
`global QR`	0.071	0.050	0.036	0.025	0.020	0.017

4.2 Distributed confidence construction

In terms of uncertainty quantification, we assess the performance of the proposed method for constructing confidence intervals by calculating the coverage probability and width of the confidence interval for each regression coefficient. For point estimation, we implement Algorithm 1 with $T = 10$ and employ the averaging-based QR estimator as the initialization. The bandwidths are set to be $h = 1.5 \cdot \{(p + \log N)/N\}^{1/3}$ and $b = 1.5 \cdot \{(p + \log n)/n\}^{1/3}$.

Table 2: Estimation error under linear and quadratic heteroscedastic models with t_2 noise, averaged over 100 trials. Results for $\tau = 0.8$, $N = nm = 150,000$ and $p = 10$, across $n = \{300, 500, 1000, 1500, 3000, 6000\}$ are reported.

Linear Heteroscedastic Model with $\tau = 0.8$, $N = 150,000$, $m = N/n$, and $p = 10$						
Methods	$n = 300$	$n = 500$	$n = 1000$	$n = 1500$	$n = 3000$	$n = 6000$
`averaging-based QR`	0.037	0.029	0.025	0.023	0.022	0.022
`distributed QR` ($T = 1$)	0.206	0.140	0.071	0.053	0.034	0.026
`distributed QR` ($T = 4$)	0.252	0.100	0.026	0.023	0.022	0.022
`distributed QR` ($T = 10$)	0.389	0.124	0.024	0.023	0.022	0.022
`distributed smoothed QR` ($T = 1$)	0.149	0.092	0.049	0.038	0.028	0.024
`distributed smoothed QR` ($T = 4$)	0.042	0.024	0.023	0.023	0.023	0.023
`distributed smoothed QR` ($T = 10$)	0.029	0.023	0.023	0.023	0.023	0.023
`global QR`	0.023	0.022	0.022	0.022	0.022	0.022
Quadratic Heteroscedastic Model with $\tau = 0.8$, $N = 150,000$, $m = N/n$, and $p = 10$						
`averaging-based QR`	0.039	0.030	0.026	0.024	0.023	0.022
`distributed QR` ($T = 1$)	0.215	0.143	0.075	0.053	0.034	0.027
`distributed QR` ($T = 4$)	0.236	0.101	0.028	0.023	0.022	0.022
`distributed QR` ($T = 10$)	0.371	0.123	0.026	0.023	0.022	0.022
`distributed smoothed QR` ($T = 1$)	0.151	0.095	0.051	0.040	0.029	0.025
`distributed smoothed QR` ($T = 4$)	0.037	0.025	0.024	0.024	0.023	0.023
`distributed smoothed QR` ($T = 10$)	0.025	0.024	0.024	0.024	0.023	0.023
`global QR`	0.023	0.023	0.023	0.023	0.022	0.022

Table 3: Estimation error under linear and quadratic heteroscedastic models with t_2 noise, averaged over 100 trials. Results for $\tau = 0.8$, $n = 300$, $p = 10$, $m = 400$, with initial estimator computed using different sample size, $n_{\text{init}} = \{150, 300, 500, 1000, 5000\}$, are reported.

Linear Heteroscedastic Model with $\tau = 0.8$, $n = 300$, $p = 10$, and $m = 400$					
Methods	$n_{\text{init}} = 150$	$n_{\text{init}} = 300$	$n_{\text{init}} = 500$	$n_{\text{init}} = 1000$	$n_{\text{init}} = 5000$
`distributed smoothed QR` ($T = 1$)	0.211	0.181	0.151	0.108	0.085
`distributed smoothed QR` ($T = 4$)	0.041	0.042	0.032	0.033	0.031
`distributed smoothed QR` ($T = 10$)	0.027	0.038	0.027	0.026	0.027
`global QR`	0.025	0.025	0.025	0.024	0.025
Quadratic Heteroscedastic Model with $\tau = 0.8$, $n = 300$, $p = 10$, and $m = 400$					
`distributed smoothed QR` ($T = 1$)	0.214	0.151	0.112	0.087	0.047
`distributed smoothed QR` ($T = 4$)	0.039	0.033	0.032	0.032	0.028
`distributed smoothed QR` ($T = 10$)	0.027	0.027	0.029	0.027	0.027
`global QR`	0.025	0.025	0.025	0.025	0.025

For confidence construction, we first consider four methods: the asymptotic normal-based interval (28) for the proposed communication-efficient estimator (CE-Normal), the normal-based interval (28) with $\widetilde{\boldsymbol{\beta}}$ replaced by $\widehat{\boldsymbol{\beta}}^{\mathrm{dc}}$ of equation (20) (DC-Normal), and the two communication-efficient bootstrap constructions as in (34) (CE-Boot (a)) and in (35) (CE-Boot (b)).

Recall that the normal-based method requires estimating the asymptotic variances σ_j^2, and the bootstrap methods depend on $\mathbf{H}^{-1}$. We consider two types of variance estimators. The first one is easier to implement but relies on the assumption that $\mathbf{H} = f_\varepsilon(0) \cdot \Sigma$, which holds when ε is independent of $\boldsymbol{x}$. In this case, $\sigma_j^2 = \tau(1-\tau)\{f_\varepsilon(0)\}^{-2}(\Sigma^{-1})_{jj}$. We compute the global density estimator $\widehat{f}_\varepsilon(0)$ as in (27) with a rule-of-thumb bandwidth $b_N^{\mathrm{rot}} = N^{-1/3} \cdot \Phi^{-1}(1-\alpha/2)^{2/3}[\{1.5 \cdot \phi(\Phi^{-1}(\tau))^2\}/\{2\Phi^{-1}(\tau)^2+1\}]^{1/3}$, and a local covariance matrix estimator $\widehat{\Sigma}_1$. The second estimator is more general and takes the form $\widehat{\sigma}_j = \{\tau(1-\tau)\}^{1/2}\left(\widehat{\mathbf{H}}_{1,b}^{-1}\widehat{\Sigma}_1\widehat{\mathbf{H}}_{1,b}^{-1}\right)_{jj}^{1/2}$, where $\widehat{\mathbf{H}}_{1,b}$ is given in (23). We generate the design matrix the same way as in Section 4.1, and focus on the following linear heteroscedastic models with different levels of heterogeneity:

$$y_i = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^* + (0.2x_{ip}+1)\{\varepsilon_i - F_{\varepsilon_i}^{-1}(\tau)\}; \tag{52}$$

$$y_i = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^* + (0.4x_{ip}+1)\{\varepsilon_i - F_{\varepsilon_i}^{-1}(\tau)\}. \tag{53}$$

We set $p = 50$, $n = 2000$, $\tau = 0.4$ and let m vary from 20 to 400. The results for 95% confidence intervals are reported in Figures 2 and 3.

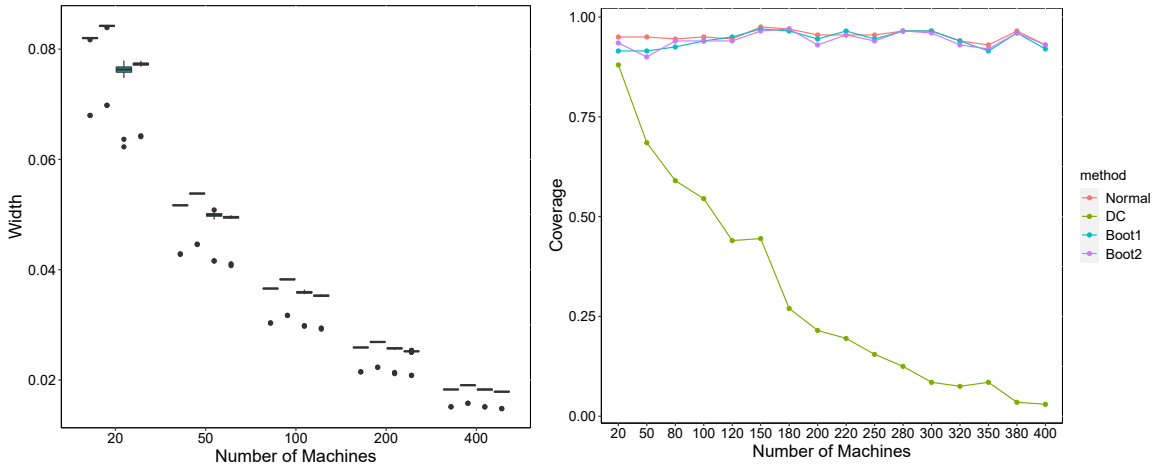
For all of our numerical results, we found that the coverage probabilities and widths of the 95% confidence intervals for the first $p-1$ regression coefficients (independent of the random noise) are similar across all methods. Specifically, the proposed CE-Normal and CE-Boot (a) & (b) methods perform very well across various model settings. Since the results across all methods are similar, they are omitted due to limited space.

We focus on reporting the empirical coverage probabilities and widths of the 95% confidence intervals for the last regression coefficient in Figures 2 and 3. We use the first type of variance estimators in the top panels and the second type in the bottom panels. From panels (b) and (d) in Figures 2 and 3, we see that the normal-based method for the simple averaging estimator suffers from severe undercoverage when the heterogeneous covariate effect is strong, which in our case, comes from the last covariate.

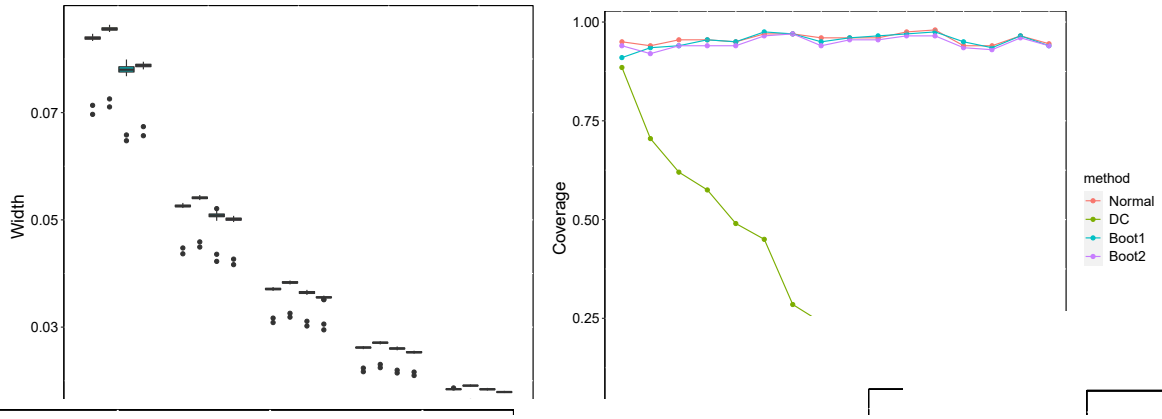
Score-based confidence sets, while computationally more intensive due to inversion of the test, are extremely efficient due to the linearity of the self-normalized representation exploited in our construction. We illustrate the improvements in a smaller scale simulation study in Section E of the Appendix.

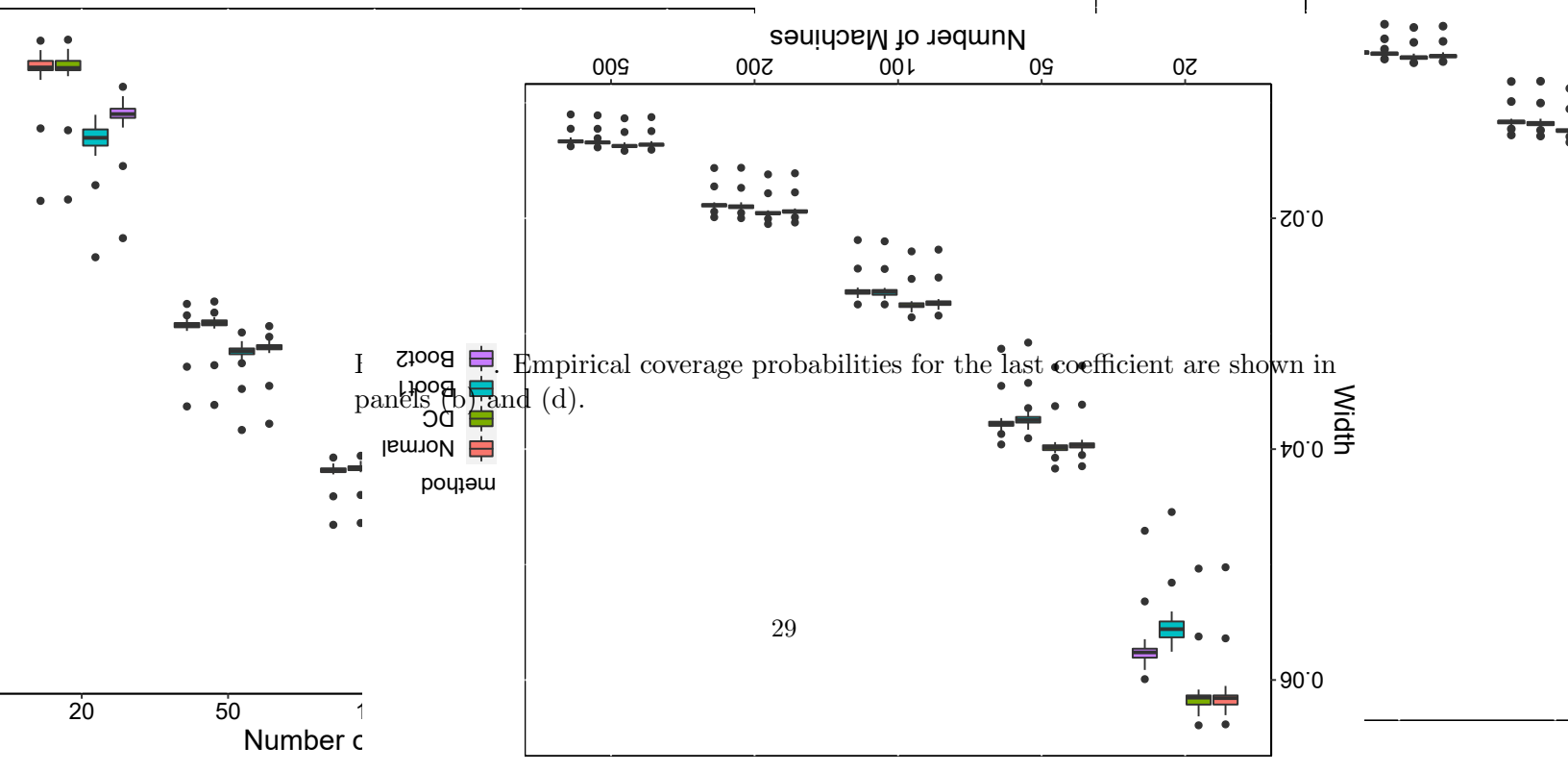
4.3 Distributed penalized quantile regression

The following numerical study illustrates the performance of the procedure proposed in Section 3, when the dimension p is larger than n for each of the m sources. For comparison purposes, we also consider the ℓ_1-penalized conquer (ℓ_1-conquer) fitted to all $N = nm$ observations, which is practically infeasible for the problems that motivated our work, and the simple averaging estimator–the average of m local ℓ_1-conquer estimates. The performance of the proposed procedure is shown for $T = 1$ and for T chosen adaptively using the



(a) (b)





I . Empirical coverage probabilities for the last coefficient are shown in panels (b) and (d).

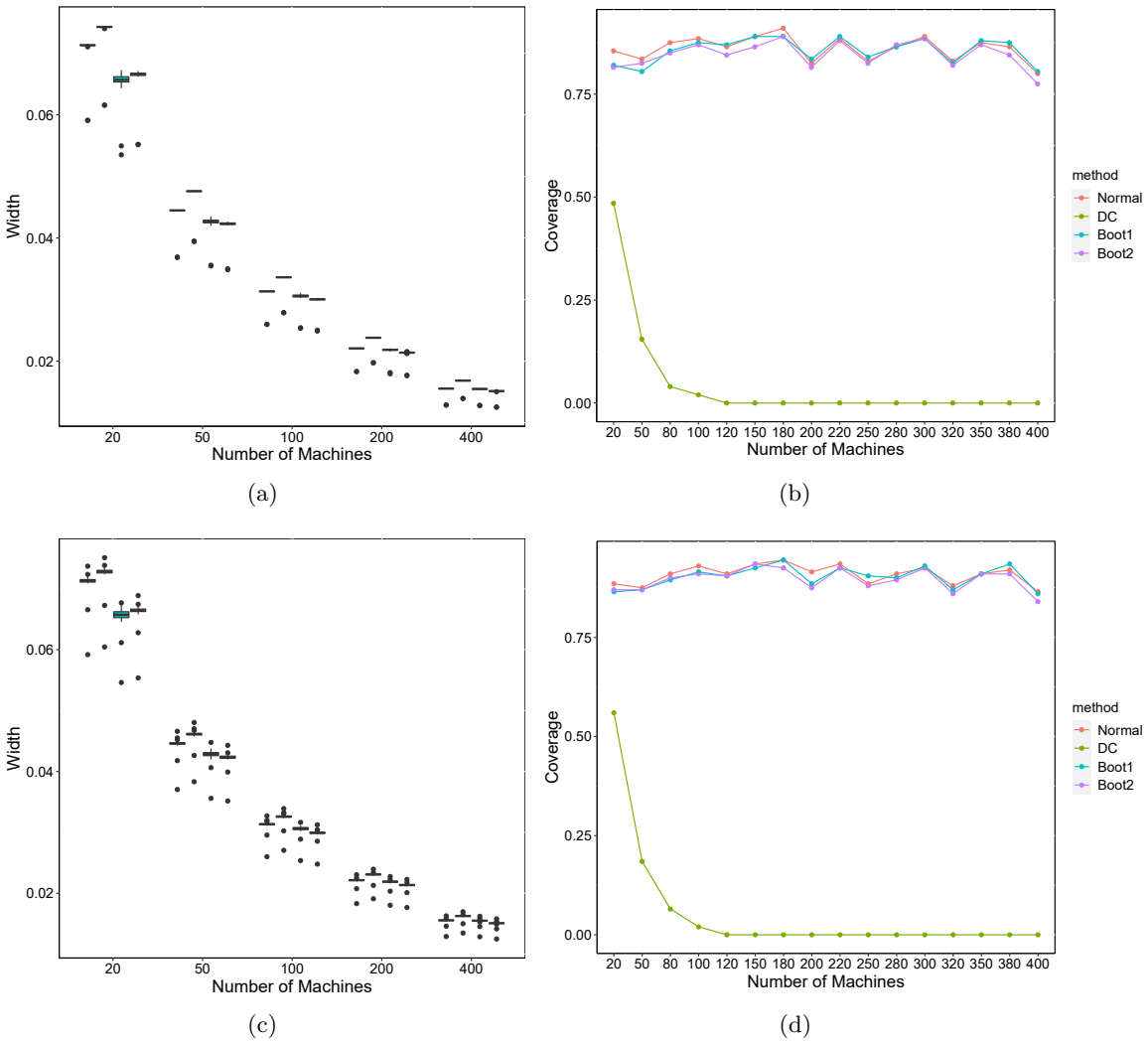


Figure 3: Confidence intervals for the regression coefficients under model (53). Other details are the same as Figure 2.

stopping criterion in Algorithm 1. As previously discussed, we use the Gaussian kernel for smoothing and the simple averaging estimator as the initialization.

We consider the following heteroscedastic models with $s = 5$ significant variables:

1. Linear heteroscedasticity: $y_i = 3 + \sum_{j=1}^{5} x_{ij} + (0.2x_{i1} + 1)\{\varepsilon_i - F_{\varepsilon_i}^{-1}(\tau)\}$;

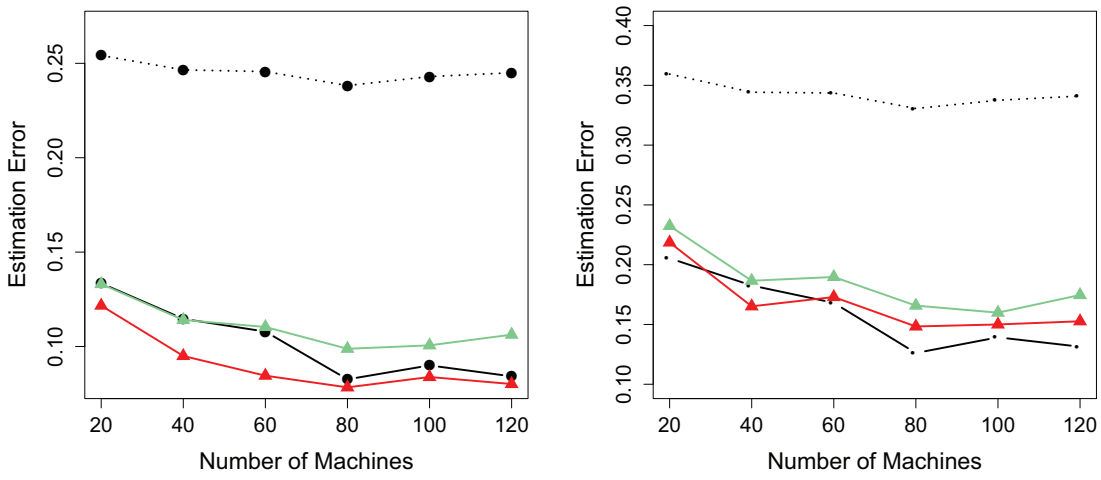
2. Quadratic heteroscedasticity: $y_i = 3 + \sum_{j=1}^{5} x_{ij} + 0.5\{1 + (0.25x_{ip} - 1)^2\}\{\varepsilon_i - F_{\varepsilon_i}^{-1}(\tau)\}$,

where $\boldsymbol{x}_i$ and ε_i are generated the same way as in Section 4.1. Moreover, we set $\tau = 0.8$, $p = 500$, $n = 400$ and $m \in \{20, 40, 60, 80, 100, 120\}$.

Guided by the theoretical results in Section 3, we set the bandwidths (b, h) as $b = 0.75s^{1/2}\{\log(p)/n\}^{1/4}$ and $h = 0.75\{s\log(p)/N\}^{1/4}$. The regularization parameter $\lambda > 0$ is selected using a validation set of size $n = 200$ for easier illustration and comparison. Note that Wang *et al.* (2017) use 60% of data for training, 20% as held-out validation set for tuning the parameters, and the remaining 20% for testing. Figure 4 provides plots of the statistical error versus number of machines, averaged over 100 Monte Carlo replications, for the proposed distributed estimator and the simple averaging estimator. The latter performs poorly under both heteroscedastic models. This is not surprising because ℓ_1-penalization induces visible finite-sample bias into the estimates, which is unaffected by aggregation no matter how many machines are available. Using this estimate as an initial value, the multi-round procedure considerably reduces the estimation error after one round of communication ($T = 1$), and eventually performs almost as well as the global ℓ_1-conquer when T is automatically determined by the stopping criterion. Since λ is tuned the same way for all the three methods, the global ℓ_1-conquer estimator does not necessarily have the best performance but still provides a yardstick for distributed estimators.

Acknowledgments

We sincerely thank the Action Editor and two anonymous reviewers for their constructive comments that help improve the previous version of the manuscript. K. M. Tan was supported by NSF Grant DMS-1949730 and DMS-2113356. H. Battey was supported by the EPSRC Fellowship EP/T01864X/1. W.-X. Zhou acknowledges the support of the NSF Grant DMS-2113409.



(a) (b)

Figure 4: Estimation error as a function of m under linear (panel (a)) and quadratic (panel (b)) heteroscedastic models with $t_{1.5}$ noise. Each point corresponds to the average of 100 Monte Carlo replications for $(n, p) = (400, 500)$. Three methods are implemented: (i) the multi-round method with $T = 1$ (—▲—▲—) and with $T = 10$ as in Algorithm 1 (—▲———▲—); (ii) the global ℓ_1-conquer estimator (—●———●—); (iii) the simple averaging estimator (·●- -●-).

Appendix A. Optimization Algorithms

A.1 First-order algorithm for solving (9)

Given an initial estimator $\widetilde{\boldsymbol{\beta}}^{(0)}$ of $\boldsymbol{\beta}^*$, and the global and local bandwidths $b, h > 0$, recall from (8) that the shifted conquer loss function takes the form $\widetilde{\mathcal{Q}}(\boldsymbol{\beta}) = \widehat{\mathcal{Q}}_{1,b}(\boldsymbol{\beta}) - \langle \nabla \widehat{\mathcal{Q}}_{1,b}(\widetilde{\boldsymbol{\beta}}^{(0)}) - \nabla \widehat{\mathcal{Q}}_h(\widetilde{\boldsymbol{\beta}}^{(0)}), \boldsymbol{\beta} \rangle$. The communication-efficient procedure involves repeatedly minimizing $\boldsymbol{\beta} \mapsto \widetilde{\mathcal{Q}}(\boldsymbol{\beta})$. Since $\widetilde{\mathcal{Q}}(\boldsymbol{\beta})$ is smooth and convex, we employ the gradient descent (GD) method which, at the kth iteration, computes

$$\widehat{\boldsymbol{\beta}}^{k+1} = \widehat{\boldsymbol{\beta}}^k - \eta_k \cdot \nabla \widetilde{\mathcal{Q}}(\widehat{\boldsymbol{\beta}}^k), \tag{54}$$

where $\eta_k > 0$ is the stepsize. The choice of stepsize is an important aspect of GD for achieving fast convergence, and has been thoroughly studied in the optimization literature. Alternatively, one can set the stepsize to be the inverse Hessian, namely, $\{\nabla^2 \widetilde{\mathcal{Q}}(\widehat{\boldsymbol{\beta}}^k)\}^{-1}$, which leads to the Newton-Raphson method. The Newton step is computationally expensive at each iteration when p is large. Moreover, when the quantile level τ is close to 0 or 1, $\nabla^2 \widetilde{\mathcal{Q}}(\widehat{\boldsymbol{\beta}}^k)$ may have a large condition number, thus causing unstableness in computing its inverse.

Motivated by the gradient-based method proposed by He *et al.* (2021) for solving (5), we consider the use of Barzilai-Borwein stepsize in (54) (Barzilai and Borwein, 1988). The main idea of the Barzilai-Borwein method is to seek a simple approximation of the inverse Hessian without having to compute it explicitly. In particular, for $k = 1, 2, \ldots$, the Barzilai-Borwein stepsizes are defined as

$$\begin{cases} \eta_{1,k} = \frac{\langle \widehat{\boldsymbol{\beta}}^k - \widehat{\boldsymbol{\beta}}^{k-1}, \widehat{\boldsymbol{\beta}}^k - \widehat{\boldsymbol{\beta}}^{k-1} \rangle}{\langle \widehat{\boldsymbol{\beta}}^k - \widehat{\boldsymbol{\beta}}^{k-1}, \nabla \widetilde{Q}(\widehat{\boldsymbol{\beta}}^k) - \nabla \widetilde{Q}(\widehat{\boldsymbol{\beta}}^{k-1}) \rangle}, \\ \eta_{2,k} = \frac{\langle \widehat{\boldsymbol{\beta}}^k - \widehat{\boldsymbol{\beta}}^{k-1}, \nabla \widetilde{Q}(\widehat{\boldsymbol{\beta}}^k) - \nabla \widetilde{Q}(\widehat{\boldsymbol{\beta}}^{k-1}) \rangle}{\langle \nabla \widetilde{Q}(\widehat{\boldsymbol{\beta}}^k) - \nabla \widetilde{Q}(\widehat{\boldsymbol{\beta}}^{k-1}), \nabla \widetilde{Q}(\widehat{\boldsymbol{\beta}}^k) - \nabla \widetilde{Q}(\widehat{\boldsymbol{\beta}}^{k-1}) \rangle}. \end{cases} \tag{55}$$

When the quantile level τ approaches 0 or 1, the objective function is flat in some directions and hence the Hessian matrix becomes more ill-conditioned. To stablize the algorithm, we set the stepsize to be $\eta_k = \min(\eta_{1,k}, \eta_{2,k}, C)$ $(k = 1, 2, \ldots)$ for some constant $C > 0$, say $C = 20$. The pseudo-code for the above Barzilai-Borwein GD method for solving (8) is given in Algorithm 2.

Algorithm 2 Gradient descent with Barzilai-Borwein stepsize (GD-BB) for solving (8).

Input: Local data vectors $\{(y_i, \boldsymbol{x}_i)\}_{i \in \mathcal{I}_1}$, $\tau \in (0,1)$, bandwidth $b \in (0,1)$, initialization $\widehat{\boldsymbol{\beta}}^0 = \widetilde{\boldsymbol{\beta}}^{(0)}$, gradient vectors $\nabla \widehat{\mathcal{Q}}_{1,b}(\widetilde{\boldsymbol{\beta}}^{(0)})$ and $\nabla \widehat{\mathcal{Q}}_h(\widetilde{\boldsymbol{\beta}}^{(0)})$, and convergence tolerance δ.

1: Compute $\widehat{\boldsymbol{\beta}}^1 \leftarrow \widehat{\boldsymbol{\beta}}^0 - \nabla \widetilde{Q}(\widehat{\boldsymbol{\beta}}^0)$.
2: **for** $k = 1, 2 \ldots$ **do**
3: Compute step sizes $\eta_{1,k}$ and $\eta_{2,k}$ defined in (55).
4: Set $\eta_k \leftarrow \min\{\eta_{1,k}, \eta_{2,k}, 20\}$ if $\eta_{1,k}, \eta_{2,k} > 0$, and $\eta_k \leftarrow 1$ otherwise.
5: Update $\widehat{\boldsymbol{\beta}}^{k+1} \leftarrow \widehat{\boldsymbol{\beta}}^k - \eta_k \nabla \widetilde{\mathcal{Q}}(\widehat{\boldsymbol{\beta}}^k)$.
6: **end for** when $\|\nabla \widetilde{\mathcal{Q}}(\widehat{\boldsymbol{\beta}}^k)\|_2 \le \delta$.

A.2 Local adaptive majorize-minimize algorithm for solving (39)

In this section, we provide an algorithm to solve the ℓ_1-penalized shifted conquer loss minimization. In particular, given an initial estimator $\widetilde{\boldsymbol{\beta}}^{(0)}$, the ℓ_1-penalized shifted conquer loss takes the form

$$\widetilde{\mathcal{Q}}(\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}_-\|_1, \text{ where } \widetilde{\mathcal{Q}}(\boldsymbol{\beta}) = \widehat{\mathcal{Q}}_{1,b}(\boldsymbol{\beta}) - \langle \nabla\widehat{\mathcal{Q}}_{1,b}(\widetilde{\boldsymbol{\beta}}^{(0)}) - \nabla\widehat{\mathcal{Q}}_h(\widetilde{\boldsymbol{\beta}}^{(0)}), \boldsymbol{\beta}\rangle. \tag{56}$$

Here $\boldsymbol{\beta}_- \in \mathbb{R}^{p-1}$ denotes the subvector of $\boldsymbol{\beta} \in \mathbb{R}^p$ with its first coordinate removed. Due to the non-differentiability of the ℓ_1-norm, the GD-BB method in Algorithm 2 is no longer applicable. By extending the majorize-minimize (MM) algorithm (Hunter and Lange, 2000) for standard quantile regressions, we employ a local adaptive majorize-minimize (LAMM) principle (Fan *et al.*, 2018) to minimize the penalized conquer loss $\boldsymbol{\beta} \mapsto \widetilde{\mathcal{Q}}(\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}_-\|_1$.

At the kth iteration with a previous estimate $\widehat{\boldsymbol{\beta}}^{k-1}$, the main idea of the LAMM algorithm is to construct an isotropic quadratic function that locally majorizes the shifted conquer loss function $\widetilde{\mathcal{Q}}(\cdot)$. Specifically, for some quadratic parameter $\phi_k > 0$, we define the quadratic function

$$F(\boldsymbol{\beta}; \phi^k, \widehat{\boldsymbol{\beta}}^{k-1}) = \widetilde{\mathcal{Q}}(\widehat{\boldsymbol{\beta}}^{k-1}) + \langle \nabla\widetilde{\mathcal{Q}}(\widehat{\boldsymbol{\beta}}^{k-1}), \boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}^{k-1}\rangle + \frac{\phi_k}{2}\|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}^{k-1}\|_2^2,$$

and then compute the update $\widehat{\boldsymbol{\beta}}^k$ by solving

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\text{minimize}} \left\{F(\boldsymbol{\beta}; \phi_k, \widehat{\boldsymbol{\beta}}^{k-1}) + \lambda\|\boldsymbol{\beta}_-\|_1\right\}. \tag{57}$$

The isotropic form of $F(\boldsymbol{\beta}; \phi_k, \widehat{\boldsymbol{\beta}}^{k-1})$, as a function of $\boldsymbol{\beta}$, permits a simple analytic solution $\widehat{\boldsymbol{\beta}}^k = (\widehat{\beta}_1^k, \ldots, \widehat{\beta}_p^k)^{\mathrm{T}}$ that takes the form

$$\begin{cases} \widehat{\beta}_1^k = \widehat{\beta}_1^{k-1} - \phi_k^{-1}\nabla_{\beta_1}\widetilde{\mathcal{Q}}(\widehat{\boldsymbol{\beta}}^{k-1}), \\ \widehat{\beta}_j^k = S(\widehat{\beta}_j^{k-1} - \phi_k^{-1}\nabla_{\beta_j}\widetilde{\mathcal{Q}}(\widehat{\boldsymbol{\beta}}^{k-1}),\ \phi_k^{-1}\lambda) \ \text{ for } j = 2, \ldots, p, \end{cases}$$

where $S(a, b) = \text{sign}(a) \cdot \max(|a| - b, 0)$ is the soft-thresholding operator. To enforce overall descent of the function value, we need $\phi_k > 0$ to be sufficiently large so that $F(\widehat{\boldsymbol{\beta}}^k; \phi_k, \widehat{\boldsymbol{\beta}}^{k-1}) \geq \widetilde{\mathcal{Q}}(\widehat{\boldsymbol{\beta}}^k)$, and hence

$$\begin{aligned} \widetilde{\mathcal{Q}}(\widehat{\boldsymbol{\beta}}^k) + \lambda\|\widehat{\boldsymbol{\beta}}_-^k\|_1 &\leq F(\widehat{\boldsymbol{\beta}}^k; \phi_k, \widehat{\boldsymbol{\beta}}^{k-1}) + \lambda\|\widehat{\boldsymbol{\beta}}_-^k\|_1 \\ &\leq F(\widehat{\boldsymbol{\beta}}^{k-1}; \phi_k, \widehat{\boldsymbol{\beta}}^{k-1}) + \lambda\|\widehat{\boldsymbol{\beta}}_-^{k-1}\|_1 = \widetilde{\mathcal{Q}}(\widehat{\boldsymbol{\beta}}^{k-1}) + \lambda\|\widehat{\boldsymbol{\beta}}_-^{k-1}\|_1. \end{aligned}$$

To choose a proper ϕ_k in practice, we start from a relatively small value $\phi_{k,0} = 0.001$, and successively inflate it by a factor $\gamma = 1.1$—$\phi_{k,\ell} = \gamma\phi_{k,\ell-1}$ for $\ell = 1, 2, \ldots$—until the majorization requirement is met. See Algorithm 3 for the pseudo-code of the LAMM algorithm for solving (39).

Appendix B. Proof of Main Results

B.1 Supporting lemmas

Recall that the data vector $(y, \boldsymbol{x}) \in \mathbb{R}\times\mathbb{R}^p$ satisfies the conditional quantile model $Q_\tau(y|\boldsymbol{x}) = F_{y|\boldsymbol{x}}^{-1}(\tau) = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*$. Equivalently, $y = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^* + \varepsilon$ with $Q_\tau(\varepsilon|\boldsymbol{x}) = 0$. Under Condition (C2),

Algorithm 3 Local adaptive majorize-minimize (LAMM) algorithm for solving (39).

Input: Local data vectors $\{(y_i, \boldsymbol{x}_i)\}_{i\in\mathcal{I}_1}$, $\tau \in (0,1)$, bandwidth $b \in (0,1)$, initialization $\widehat{\boldsymbol{\beta}}^0 = \widetilde{\boldsymbol{\beta}}^{(0)}$, gradient vectors $\nabla\widehat{\mathcal{Q}}_{1,b}(\widetilde{\boldsymbol{\beta}}^{(0)})$ and $\nabla\widehat{\mathcal{Q}}_h(\widetilde{\boldsymbol{\beta}}^{(0)})$, regularization parameter $\lambda > 0$, isotropic parameter ϕ_0 and convergence tolerance δ.

1: **for** $k = 1, 2 \ldots$ **do**
2: Set $\phi_k \leftarrow \max\{\phi_0, \phi_{k-1}/1.1\}$.
3: **repeat**
4: Update $\widehat{\beta}_1^k \leftarrow \widehat{\beta}_1^{k-1} - \phi_k^{-1}\nabla_{\beta_1}\widetilde{\mathcal{Q}}(\widehat{\boldsymbol{\beta}}^{k-1})$.
5: Update $\widehat{\beta}_j^k \leftarrow S(\widehat{\beta}_j^{k-1} - \phi_k^{-1}\nabla_{\beta_j}\widetilde{\mathcal{Q}}(\widehat{\boldsymbol{\beta}}^{k-1}),\ \phi_k^{-1}\lambda)$ for $j = 2, \ldots, p$.
6: **If** $F(\widehat{\boldsymbol{\beta}}^k; \phi_k, \widehat{\boldsymbol{\beta}}^{k-1}) < \widetilde{\mathcal{Q}}(\widehat{\boldsymbol{\beta}}^k)$, set $\phi_k \leftarrow 1.1\phi_k$.
7: **until** $F(\widehat{\boldsymbol{\beta}}^k; \phi_k, \widehat{\boldsymbol{\beta}}^{k-1}) \geq \widetilde{\mathcal{Q}}(\widehat{\boldsymbol{\beta}}^k)$.
8: **end for** when $\|\widehat{\boldsymbol{\beta}}^k - \widehat{\boldsymbol{\beta}}^{k-1}\|_2 \leq \delta$.

$\boldsymbol{z} = \Sigma^{-1/2}\boldsymbol{x} \in \mathbb{R}^p$ denotes the standardized vector of covariates such that $\mathbb{E}(\boldsymbol{z}\boldsymbol{z}^{\mathrm{T}}) = \mathbf{I}_p$. For every $\delta \in (0,1]$, define

$$\gamma_\delta = \inf\big\{\gamma > 0 : \sup_{\boldsymbol{u}\in\mathbb{S}^{p-1}} \mathbb{E}\big\{(\boldsymbol{z}^{\mathrm{T}}\boldsymbol{u})^2 \mathbb{1}(|\boldsymbol{z}^{\mathrm{T}}\boldsymbol{u}| > \gamma_\delta)\big\} \leq \delta\big\}. \tag{58}$$

By the sub-Gaussian assumption on $\boldsymbol{x}$, γ_δ depends only on δ and υ_1, and the map $\delta \mapsto \gamma_\delta$ is non-increasing with $\gamma_\delta \downarrow 0$ as $\delta \to 1$. Under a weaker condition that $\boldsymbol{z}$ has *uniformly bounded fourth moments*, namely, $\mu_4 = \sup_{\boldsymbol{u}\in\mathbb{S}^{p-1}} \mathbb{E}(\boldsymbol{z}^{\mathrm{T}}\boldsymbol{u})^4 < \infty$, we have $\gamma_\delta \leq (\mu_4/\delta)^{1/2}$.

For the conditional density $f_{\varepsilon|\boldsymbol{x}}(\cdot)$, Condition (C1) ensures that as long as b is sufficiently small,

$$\underline{f}_b \leq \min_{|u|\leq b/2} f_{\varepsilon|\boldsymbol{x}}(u) \leq \max_{|u|\leq b/2} f_{\varepsilon|\boldsymbol{x}}(u) \leq \bar{f}_b \ \text{ almost surely (over } \boldsymbol{x})$$

for some constants $\bar{f}_b \geq \underline{f}_b > 0$. For example, we may take $\underline{f}_b = \underline{f} - l_0 b/2$ and $\bar{f}_b = \bar{f} + l_0 b/2$. Given a kernel $K(\cdot)$ and bandwidth $b > 0$, recall that $\widehat{\mathcal{Q}}_{1,b}(\boldsymbol{\beta}) = (1/n)\sum_{i\in\mathcal{I}_1} \ell_b(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})$ denotes the local smoothed loss function, where $\ell_b(\cdot) = (\rho_\tau * K_b)(\cdot)$. The proof of Theorem 1 depends heavily on the local strong convexity of $\widehat{\mathcal{Q}}_{1,b}(\cdot)$ in a neighborhood of $\boldsymbol{\beta}^*$. To this end, we first introduce the notion of symmetrized Bregman divergence. For any differentiable convex function $\psi : \mathbb{R}^k \to \mathbb{R}$ ($k \geq 1$), the corresponding Bregman divergence is given by $D_\psi(\boldsymbol{w}', \boldsymbol{w}) = \psi(\boldsymbol{w}') - \psi(\boldsymbol{w}) - \langle\nabla\psi(\boldsymbol{w}), \boldsymbol{w}' - \boldsymbol{w}\rangle$. Define its symmetrized version as

$$\overline{D}_\psi(\boldsymbol{w}, \boldsymbol{w}') = D_\psi(\boldsymbol{w}, \boldsymbol{w}') + D_\psi(\boldsymbol{w}', \boldsymbol{w}) = \big\langle\nabla\psi(\boldsymbol{w}) - \nabla\psi(\boldsymbol{w}'), \boldsymbol{w} - \boldsymbol{w}'\big\rangle, \quad \boldsymbol{w}, \boldsymbol{w}' \in \mathbb{R}^k. \tag{59}$$

The first two lemmas, Lemma 15 and 16, provide a lower bound on the symmetrized Bregman divergence of the shifted conquer loss $\widetilde{\mathcal{Q}}(\cdot)$ given in (8) and an upper bound on the global gradient, respectively. These two results coincide with Lemmas A.1 and A.2 in the supplement of He *et al.* (2021). We reproduce them here for the sake of readability. In particular, the former implies the restricted strong convexity property of $\widetilde{\mathcal{Q}}(\cdot)$.

Lemma 15 *For any $x > 0$ and $0 < r \leq b/(4\gamma_{0.25})$,*

$$\inf_{\boldsymbol{\beta}\in\Theta(r)} \frac{\overline{D}_{\widetilde{\mathcal{Q}}}(\boldsymbol{\beta}, \boldsymbol{\beta}^*)}{\kappa_l\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_\Sigma^2} \geq \frac{3}{4}\underline{f}_b - \bar{f}_b^{1/2}\left(\frac{5}{4}\sqrt{\frac{bp}{r^2 n}} + \sqrt{\frac{bx}{8r^2 n}}\right) - \frac{bx}{3r^2 n} \tag{60}$$

with probability at least $1 - e^{-x}$, where $\kappa_l = \min_{|u|\leq 1} K(u) > 0$.

Consider the gradient $\nabla\widehat{\mathcal{Q}}_h(\cdot)$ evaluated at $\boldsymbol{\beta}^*$, namely,

$$\nabla\widehat{\mathcal{Q}}_h(\boldsymbol{\beta}^*) = \frac{1}{N}\sum_{i=1}^{N}\big\{\overline{K}(-\varepsilon_i/h) - \tau\big\}\boldsymbol{x}_i,$$

where $\varepsilon_i = y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*$. The following lemma provides an upper bound on the ℓ_2-norm of $\nabla\widehat{\mathcal{Q}}_h(\boldsymbol{\beta}^*)$. Recall that $\Omega = \Sigma^{-1}$, and we write $\|\boldsymbol{u}\|_\Omega = \|\Sigma^{-1/2}\boldsymbol{u}\|_2$ for $\boldsymbol{u}\in\mathbb{R}^p$.

Lemma 16 *Conditions (C1)–(C3) ensure that, for any $x > 0$,*

$$\|\nabla\widehat{\mathcal{Q}}_h(\boldsymbol{\beta}^*)\|_\Omega \le C_0\bigg(\sqrt{\frac{p+x}{N}} + h^2\bigg) \tag{61}$$

with probability at least $1-e^{-x}$ as long as $N \gtrsim p+x$, where $C_0 > 0$ is a constant depending only on $(\tau, l_0, \upsilon_1, \kappa_2)$.

Next, we extend the above results to the high-dimensional setting in which $p \gg n$. Recall that $\Theta(r)$ $(r>0)$ denotes the local ℓ_2 neighborhood of $\boldsymbol{\beta}^*$ under $\|\cdot\|_\Sigma$-norm. Furthermore, define the ℓ_1-cone Λ as in (40), that is,

$$\Lambda = \big\{\boldsymbol{\beta}\in\mathbb{R}^p : \|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_1 \le 4s^{1/2}\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_\Sigma\big\}. \tag{62}$$

Lemma 17 *Assume Conditions (C1), (C2) and (C4) hold. Then, for any $x>0$, $0 < r \le b/(4\gamma_{0.25})$ and $L>0$,*

$$\inf_{\boldsymbol{\beta}\in\Theta(r)\cap\Lambda}\frac{\overline{D}_{\widetilde{\mathcal{Q}}}(\boldsymbol{\beta},\boldsymbol{\beta}^*)}{\kappa_l\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_\Sigma^2} \ge \frac{3}{4}\underline{f}_b - \bar{f}_b^{1/2}\left\{5B\sqrt{\frac{2bs\log(2p)}{r^2n}} + \sqrt{\frac{bx}{8r^2n}}\right\} - \frac{bx}{3r^2n} \tag{63}$$

with probability at least $1-e^{-x}$.

Proof of Lemma 17: Following the proof of Lemma 4.1 in Tan, Wang and Zhou (2022), it suffices to bound

$$\mathbb{E}\bigg\|\frac{1}{n}\sum_{i=1}^{n}e_i\psi_{b/2}(\varepsilon_i)\boldsymbol{x}_i\bigg\|_\infty = \mathbb{E}\,\mathbb{E}_e\bigg\|\frac{1}{n}\sum_{i=1}^{n}e_i\psi_{b/2}(\varepsilon_i)\boldsymbol{x}_i\bigg\|_\infty, \tag{64}$$

where $e_1,\ldots,e_n$ are i.i.d. Rademacher random variables, $\psi_{b/2}(\varepsilon_i) = \mathbb{1}(|\varepsilon_i|\le b/2)$, and $\mathbb{E}_e$ denotes the (conditional) expectation over $e_1,\ldots,e_n$ given all the remaining random variables. Applying Hoeffding's moment inequality yields

$$\begin{aligned}&\mathbb{E}_e\bigg\|\frac{1}{n}\sum_{i=1}^{n}e_i\psi_{b/2}(\varepsilon_i)\boldsymbol{x}_i\bigg\|_\infty\\ &\le \max_{1\le j\le p}\left\{\frac{1}{n}\sum_{i=1}^{n}x_{ij}^2\psi_{b/2}^2(\varepsilon_i)\right\}^{1/2}\sqrt{\frac{2\log(2p)}{n}} \le \left\{\frac{1}{n}\sum_{i=1}^{n}\psi_{b/2}(\varepsilon_i)\right\}^{1/2}B\sqrt{\frac{2\log(2p)}{n}}.\end{aligned}$$

Moreover, note that $\mathbb{E}\{\psi_{b/2}(\varepsilon_i)|\boldsymbol{x}_i\} \le \bar{f}_b \cdot b$. Substituting these bounds into (64), we obtain that

$$\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n} e_i\psi_{b/2}(\varepsilon_i)\boldsymbol{x}_i\right\|_\infty \le (2\bar{f}_b)^{1/2}B\sqrt{\frac{b\log(2p)}{n}}.$$

Keep the rest of the proof the same, we obtain the claimed bound (63).

For the empirical loss $\widehat{\mathcal{Q}}_h(\boldsymbol{\beta}) = (1/N)\sum_{i=1}^{N}\ell_h(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})$, define its population counterpart $\mathcal{Q}_h(\boldsymbol{\beta}) = \mathbb{E}\widehat{\mathcal{Q}}_h(\boldsymbol{\beta}) = \mathbb{E}_{(y,\boldsymbol{x})\sim P}\{\ell_h(y - \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta})\}$.

Lemma 18 *Conditions (C1), (C2) and (C4) ensure that, for any $x > 0$,*

$$\|\nabla\widehat{\mathcal{Q}}_h(\boldsymbol{\beta}^*) - \nabla\mathcal{Q}_h(\boldsymbol{\beta}^*)\|_\infty \le C(\tau,h)\sqrt{\frac{\log(2p)+x}{N}} + B\max(\tau,1-\tau)\frac{\log(2p)+x}{3N}$$

with probability at least $1 - e^{-x}$, where $C(\tau,h) = \sigma_u\sqrt{2\{\tau(1-\tau)+(1+\tau)l_0\kappa_2h^2\}}$ and $\sigma_u = \max_{1\le j\le p}\sigma_{jj}^{1/2}$.

Proof of Lemma 18: To begin with, write

$$\|\nabla\mathcal{Q}_h(\boldsymbol{\beta}^*) - \nabla\mathcal{Q}_h(\boldsymbol{\beta}^*)\|_\infty = \max_{1\le j\le p}\left|\frac{1}{N}\sum_{i=1}^{N}(1-\mathbb{E})w_ix_{ij}\right|,$$

where $w_i = \overline{K}(-\varepsilon_i/h) - \tau$ for $i = 1,\ldots,n$, and note that $|w_ix_{ij}| \le B\bar{\tau}$ with $\bar{\tau} := \max(\tau, 1-\tau)$. Using Taylor series expansion and integration by parts, we obtain that

$$|\mathbb{E}\big(w_i|\boldsymbol{x}_i\big)| \le 0.5l_0\kappa_2h^2 \quad\text{and}\quad \mathbb{E}\big\{\overline{K}^2(-\varepsilon_i/h)|\boldsymbol{x}_i\big\} \le \tau + l_0\kappa^2h^2,$$

which in turn implies $\mathbb{E}(w_i^2|\boldsymbol{x}_i) \le \tau(1-\tau)+(1+\tau)l_0\kappa_2h^2 = \tau(1-\tau)+Ch^2$. Hence, applying Bernstein's inequality yields that, for any $1 \le j \le p$ and $z \ge 0$,

$$\left|\frac{1}{N}\sum_{i=1}^{N}(1-\mathbb{E})w_ix_{ij}\right| \le \sigma_{jj}^{1/2}\sqrt{2\big\{\tau(1-\tau)+Ch^2\big\}\frac{z}{N}} + \frac{B\bar{\tau}}{3}\frac{z}{N}$$

with probability at least $1 - 2e^{-z}$. Taking $z = \log(2p) + x$, the claimed bound follows immediately from the union bound.

With the above preparations, we are ready to prove the main results in the paper.

B.2 Proof of Theorem 1

Proof of (11). The proof is carried out conditioning on the "good" event that $\widetilde{\boldsymbol{\beta}}^{(0)} \in \Theta(r_0)$. Let $\widetilde{\boldsymbol{\beta}} = \widetilde{\boldsymbol{\beta}}^{(1)}$ be the one-step estimator that minimizes $\widetilde{\mathcal{Q}}(\cdot)$. Set $r_{\mathrm{loc}} = b/(4\gamma_{0.25})$ with $\gamma_{0.25}$ given in (58), and define an intermediate estimator $\widetilde{\boldsymbol{\beta}}_\eta = \boldsymbol{\beta}^* + \eta(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)$ with

$$\eta = \sup\big\{u \in [0,1] : \boldsymbol{\beta}^* + u(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \in \Theta(r_{\mathrm{loc}})\big\}\begin{cases} = 1 & \text{if } \widetilde{\boldsymbol{\beta}} \in \Theta(r_{\mathrm{loc}}), \\ \in (0,1) & \text{if } \widetilde{\boldsymbol{\beta}} \notin \Theta(r_{\mathrm{loc}}). \end{cases}$$

In other words, η is the largest value of $u \in (0,1]$ such that the corresponding convex combination of $\boldsymbol{\beta}^*$ and $\widetilde{\boldsymbol{\beta}}$—namely, $(1-u)\boldsymbol{\beta}^* + u\widetilde{\boldsymbol{\beta}}$—falls into the region $\Theta(r_0)$. Hence, if $\widetilde{\boldsymbol{\beta}} \notin \Theta(r_{\mathrm{loc}})$, we must have $\widetilde{\boldsymbol{\beta}}_\eta \in \partial\Theta(r_{\mathrm{loc}}) = \{\boldsymbol{\beta} \in \mathbb{R}^p : \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_\Sigma = r_{\mathrm{loc}}\}$.

By Lemma C.1 in Sun, Zhou and Fan (2020), the three points $\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\beta}}_\eta$ and $\boldsymbol{\beta}^*$ satisfy $\overline{D}_{\widetilde{\mathcal{Q}}}(\widetilde{\boldsymbol{\beta}}_\eta, \boldsymbol{\beta}^*) \le \eta \overline{D}_{\widetilde{\mathcal{Q}}}(\widetilde{\boldsymbol{\beta}}, \boldsymbol{\beta}^*)$, where by (59),

$$\overline{D}_{\widetilde{\mathcal{Q}}}(\boldsymbol{\beta}, \boldsymbol{\beta}^*) = \big\langle \nabla\widetilde{\mathcal{Q}}(\boldsymbol{\beta}) - \nabla\widetilde{\mathcal{Q}}(\boldsymbol{\beta}^*), \boldsymbol{\beta} - \boldsymbol{\beta}^* \big\rangle = \big\langle \nabla\widehat{\mathcal{Q}}_{1,b}(\boldsymbol{\beta}) - \nabla\widehat{\mathcal{Q}}_{1,b}(\boldsymbol{\beta}^*), \boldsymbol{\beta} - \boldsymbol{\beta}^* \big\rangle = \overline{D}_{\widehat{\mathcal{Q}}_{1,b}}(\boldsymbol{\beta}, \boldsymbol{\beta}^*)$$

for $\boldsymbol{\beta} \in \mathbb{R}^p$. Taking into account the first-order optimality condition $\nabla\widetilde{\mathcal{Q}}(\widetilde{\boldsymbol{\beta}}) = \mathbf{0}$, it follows that

$$\overline{D}_{\widetilde{\mathcal{Q}}}(\widetilde{\boldsymbol{\beta}}_\eta, \boldsymbol{\beta}^*) \le -\eta \big\langle \nabla\widetilde{\mathcal{Q}}(\boldsymbol{\beta}^*), \widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \big\rangle \le \|\nabla\widetilde{\mathcal{Q}}(\boldsymbol{\beta}^*)\|_\Omega \cdot \|\widetilde{\boldsymbol{\beta}}_\eta - \boldsymbol{\beta}^*\|_\Sigma. \tag{65}$$

For the left-hand side of (65), applying Lemma 15 yields that with probability at least $1 - e^{-x}$,

$$\overline{D}_{\widetilde{\mathcal{Q}}}(\boldsymbol{\beta}, \boldsymbol{\beta}^*) \ge 0.5 \underline{f}\kappa_l \cdot \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_\Sigma^2 \tag{66}$$

holds uniformly over all $\boldsymbol{\beta} \in \Theta(r_{\mathrm{loc}})$ as long as $(p+x)/n \lesssim b \lesssim 1$.

To bound the right-hand side of (65), we define vector-valued random processes

$$\begin{cases} \Delta_1(\boldsymbol{\beta}) = \Sigma^{-1/2}\big\{\nabla\widehat{\mathcal{Q}}_{1,b}(\boldsymbol{\beta}) - \nabla\widehat{\mathcal{Q}}_{1,b}(\boldsymbol{\beta}^*) - \mathbf{H}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)\big\}, \\ \Delta(\boldsymbol{\beta}) = \Sigma^{-1/2}\big\{\nabla\widehat{\mathcal{Q}}_h(\boldsymbol{\beta}) - \nabla\widehat{\mathcal{Q}}_h(\boldsymbol{\beta}^*) - \mathbf{H}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)\big\}, \end{cases} \tag{67}$$

where $\Sigma = \mathbb{E}(\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}})$ and $\mathbf{H} = \mathbb{E}\{f_{\varepsilon|\boldsymbol{x}}(0)\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\}$. Following the proof of Theorem 4.2 in He *et al.* (2021), it can be shown that, with probability at least $1 - 2e^{-x}$,

$$\sup_{\boldsymbol{\beta}\in\Theta(r)} \|\Delta_1(\boldsymbol{\beta})\|_2 \le C_1 r\left(\sqrt{\frac{p+x}{nb}} + r + b\right) \quad \text{and} \quad \sup_{\boldsymbol{\beta}\in\Theta(r)} \|\Delta(\boldsymbol{\beta})\|_2 \le C_1 r\left(\sqrt{\frac{p+x}{Nh}} + r + h\right) \tag{68}$$

as long as $b \gtrsim \sqrt{(p+x)/n}$, $h \gtrsim \sqrt{(p+x)/N}$ and $n \gtrsim p + x$, where $C_1 > 0$ is a constant independent of (N, n, p, h, b).

Recall that $\nabla\widetilde{\mathcal{Q}}(\boldsymbol{\beta}) = \nabla\widetilde{\mathcal{Q}}_{1,b}(\boldsymbol{\beta}) - \nabla\widetilde{\mathcal{Q}}_{1,b}(\widetilde{\boldsymbol{\beta}}^{(0)}) + \nabla\widetilde{\mathcal{Q}}_h(\widetilde{\boldsymbol{\beta}}^{(0)})$ and $b \ge h$. Hence, applying (68) yields that, conditioned on the event $\mathcal{E}_0(r_0) \cap \mathcal{E}_*(r_*)$,

$$\begin{aligned} \|\nabla\widetilde{\mathcal{Q}}(\boldsymbol{\beta}^*)\|_\Omega &= \|\Delta(\widetilde{\boldsymbol{\beta}}^{(0)}) - \Delta_1(\widetilde{\boldsymbol{\beta}}^{(0)}) + \Sigma^{-1/2}\nabla\widehat{\mathcal{Q}}_h(\boldsymbol{\beta}^*)\|_2 \\ &\le \|\Delta(\widetilde{\boldsymbol{\beta}}^{(0)})\|_2 + \|\Delta_1(\widetilde{\boldsymbol{\beta}}^{(0)})\|_2 + \|\nabla\widehat{\mathcal{Q}}_h(\boldsymbol{\beta}^*)\|_\Omega \\ &\le C_1\left(\sqrt{\frac{p+x}{nb}} + \sqrt{\frac{p+x}{Nh}} + 2r_0 + 2b\right) \cdot r_0 + r_*. \end{aligned} \tag{69}$$

Together, the bounds (65), (66) and (69) imply that, conditioned on $\mathcal{E}_0(r_0) \cap \mathcal{E}_*(r_*)$,

$$\begin{aligned} &\|\widetilde{\boldsymbol{\beta}}_\eta - \boldsymbol{\beta}^*\|_\Sigma \le 2(\underline{f}\kappa_l)^{-1}\|\nabla\widetilde{\mathcal{Q}}(\boldsymbol{\beta}^*)\|_\Omega \\ &\le 2(\underline{f}\kappa_l)^{-1}\left\{C_1\left(\sqrt{\frac{p+x}{nb}} + \sqrt{\frac{p+x}{Nh}} + 2r_0 + 2b\right) \cdot r_0 + r_*\right\} \end{aligned} \tag{70}$$

with probability at least $1 - 3e^{-x}$. We further let the bandwidths $b \geq h > 0$ satisfy $b \gtrsim \max(r_0, r_*)$ and $\sqrt{(p+x)/(nb)} + b + \sqrt{(p+x)/(Nh)} \lesssim 1$, so that the right-hand of (70) is strictly less than r_{loc}. Then, the intermediate point $\widetilde{\boldsymbol{\beta}}_\eta$—a convex combination of $\boldsymbol{\beta}^*$ and $\widetilde{\boldsymbol{\beta}}$—falls into the interior of the local region $\Theta(r_0)$ with high probability conditioned on $\mathcal{E}_0(r_0) \cap \mathcal{E}_*(r_*)$. Via proof by contradiction, we must have $\widetilde{\boldsymbol{\beta}} \in \Theta(r_{\mathrm{loc}})$ and hence $\widetilde{\boldsymbol{\beta}}_\eta = \widetilde{\boldsymbol{\beta}}$; otherwise if $\widetilde{\boldsymbol{\beta}} \notin \Theta(r_{\mathrm{loc}})$, by construction $\widetilde{\boldsymbol{\beta}}_\eta$ lies on the boundary of $\Theta(r_{\mathrm{loc}})$, which is a contradiction. As a result, the bound (70) also applies to $\widetilde{\boldsymbol{\beta}}$, as desired.

Proof of (12). To establish the Bahadur representation, note that the random process $\Delta_1(\cdot)$ defined in (67) can be written as $\Delta_1(\boldsymbol{\beta}) = \Sigma^{-1/2}\{\nabla\widetilde{\mathcal{Q}}(\boldsymbol{\beta}) - \nabla\widetilde{\mathcal{Q}}(\boldsymbol{\beta}^*) - \mathbf{H}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)\}$. Moreover, note that

$$\nabla\widetilde{\mathcal{Q}}(\boldsymbol{\beta}^*) - \nabla\widehat{\mathcal{Q}}_h(\boldsymbol{\beta}^*) = \nabla\widehat{\mathcal{Q}}_{1,b}(\boldsymbol{\beta}^*) - \nabla\widehat{\mathcal{Q}}_{1,b}(\widetilde{\boldsymbol{\beta}}^{(0)}) + \nabla\widehat{\mathcal{Q}}_h(\widetilde{\boldsymbol{\beta}}^{(0)}) - \nabla\widehat{\mathcal{Q}}_h(\boldsymbol{\beta}^*),$$

which in turn implies

$$\|\nabla\widetilde{\mathcal{Q}}(\boldsymbol{\beta}^*) - \nabla\widehat{\mathcal{Q}}_h(\boldsymbol{\beta}^*)\|_\Omega \leq \|\Delta_1(\widetilde{\boldsymbol{\beta}}^{(0)})\|_2 + \|\Delta(\widetilde{\boldsymbol{\beta}}^{(0)})\|_2,$$

where $\Delta(\cdot)$ is given in (67). Recall that $\nabla\widetilde{\mathcal{Q}}(\widetilde{\boldsymbol{\beta}}) = \mathbf{0}$, and conditioned on $\mathcal{E}_0(r_0) \cap \mathcal{E}_*(r_*)$,

$$\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_\Sigma \leq r_1 \asymp \left(\sqrt{\frac{p+x}{nb}} + \sqrt{\frac{p+x}{Nh}} + b\right) \cdot r_0 + r_*$$

with high probability. Under the assumed constraints on b, h and $r_0 \gtrsim r_*$, we may assume $r_1 \leq r_0$. Consequently,

$$\begin{aligned}
&\|\nabla\widehat{\mathcal{Q}}_h(\boldsymbol{\beta}^*) + \mathbf{H}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_\Omega \\
&= \|\Sigma^{-1/2}\nabla\widehat{\mathcal{Q}}_h(\boldsymbol{\beta}^*) - \Sigma^{-1/2}\nabla\widetilde{\mathcal{Q}}(\boldsymbol{\beta}^*) - \Delta_1(\widetilde{\boldsymbol{\beta}})\|_2 \\
&\leq \|\Delta_1(\widetilde{\boldsymbol{\beta}}^{(0)})\|_2 + \|\Delta(\widetilde{\boldsymbol{\beta}}^{(0)})\|_2 + \|\Delta_1(\widetilde{\boldsymbol{\beta}})\|_2 \leq 2\sup_{\boldsymbol{\beta}\in\Theta(r_0)}\|\Delta_1(\boldsymbol{\beta})\|_2 + \sup_{\boldsymbol{\beta}\in\Theta(r_0)}\|\Delta(\boldsymbol{\beta})\|_2
\end{aligned} \tag{71}$$

with the same probability conditioned on $\mathcal{E}_0(r_0) \cap \mathcal{E}_*(r_*)$. Combining (71) with the bounds in (68) completes the proof of (12).

B.3 Proof of Theorem 2

Recall that, for $t = 1, 2, \ldots$, $\widetilde{\mathcal{Q}}^{(t)}(\cdot)$ given in (13) denotes the shifted smoothed QR loss at iteration t, whose gradient and Hessian are

$$\nabla\widetilde{\mathcal{Q}}^{(t)}(\boldsymbol{\beta}) = \nabla\widehat{\mathcal{Q}}_{1,b}(\boldsymbol{\beta}) - \nabla\widehat{\mathcal{Q}}_{1,b}(\widetilde{\boldsymbol{\beta}}^{(t-1)}) + \nabla\widehat{\mathcal{Q}}_h(\widetilde{\boldsymbol{\beta}}^{(t-1)}) \quad \text{and} \quad \nabla^2\widetilde{\mathcal{Q}}^{(t)}(\boldsymbol{\beta}) = \nabla^2\widehat{\mathcal{Q}}_{1,b}(\boldsymbol{\beta}).$$

Let $\Delta_1(\boldsymbol{\beta})$ and $\Delta(\boldsymbol{\beta})$ be the stochastic processes defined in the proof of Theorem 1. The gradient $\nabla\widetilde{\mathcal{Q}}^{(t)}(\boldsymbol{\beta}^*)$ can thus be written as $\Sigma^{-1/2}\nabla\widetilde{\mathcal{Q}}^{(t)}(\boldsymbol{\beta}^*) = \{\Delta(\widetilde{\boldsymbol{\beta}}^{(t-1)}) - \Delta_1(\widetilde{\boldsymbol{\beta}}^{(t-1)})\} + \Sigma^{-1/2}\nabla\widehat{\mathcal{Q}}_h(\boldsymbol{\beta}^*)$, so that

$$\|\nabla\widetilde{\mathcal{Q}}^{(t)}(\boldsymbol{\beta}^*)\|_\Omega \leq \|\Delta(\widetilde{\boldsymbol{\beta}}^{(t-1)})\|_2 + \|\Delta_1(\widetilde{\boldsymbol{\beta}}^{(t-1)})\|_2 + \|\nabla\widehat{\mathcal{Q}}_h(\boldsymbol{\beta}^*)\|_\Omega. \tag{72}$$

Given a sequence of iterates $\{\widetilde{\boldsymbol{\beta}}^{(t)}\}_{t=0,1,\ldots,T}$, we define the "good" events

$$\mathcal{E}_t(r_t) = \big\{\widetilde{\boldsymbol{\beta}}^{(t)} \in \Theta(r_t)\big\}, \quad t = 0, \ldots, T,$$

for some sequence of radii $r_0 \geq r_1 \geq \cdots \geq r_T > 0$ to be determined. Moreover, note that all the shifted loss functions $\widetilde{\mathcal{Q}}^{(t)}(\cdot)$ have the same symmetrized Bregman divergence, denoted by

$$\overline{D}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = \langle \nabla \widetilde{\mathcal{Q}}^{(t)}(\boldsymbol{\beta}_1) - \nabla \widetilde{\mathcal{Q}}^{(t)}(\boldsymbol{\beta}_2), \boldsymbol{\beta}_1 - \boldsymbol{\beta}_2 \rangle = \langle \nabla \widehat{\mathcal{Q}}_{1,b}(\boldsymbol{\beta}_1) - \nabla \widehat{\mathcal{Q}}_{1,b}(\boldsymbol{\beta}_2), \boldsymbol{\beta}_1 - \boldsymbol{\beta}_2 \rangle.$$

In other words, the curvature of shifted loss functions is only determined by the local loss $\widehat{\mathcal{Q}}_{1,b}(\cdot)$. Define the local radius $r_{\mathrm{loc}} = b/(4\gamma_{0.25})$ the same way as in the proof of Theorem 1. As long as $(p+x)/n \lesssim b \lesssim 1$, Lemma 15 ensures that with probability at least $1 - e^{-x}$,

$$\overline{D}(\boldsymbol{\beta}, \boldsymbol{\beta}^*) \geq 0.5 \underline{f} \kappa_l \cdot \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_{\Sigma}^2 = \kappa \cdot \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_{\Sigma}^2 \tag{73}$$

holds uniformly over all $\boldsymbol{\beta} \in \Theta(r_{\mathrm{loc}})$, where $\kappa := 0.5 \underline{f} \kappa_l$ is the curvature parameter. Let $\mathcal{E}_{\mathrm{loc}}$ be the event that the local strong convexity (73) holds.

Proceeding via proof by contradiction, at each iteration $t \geq 1$, we may construct an intermediate estimator $\widetilde{\boldsymbol{\beta}}_{\mathrm{imd}}^{(t)}$—as a convex combination of $\widetilde{\boldsymbol{\beta}}^{(t)}$ and $\boldsymbol{\beta}^*$—which falls in $\Theta(r_{\mathrm{loc}})$. If event $\mathcal{E}_*(r_*) \cap \mathcal{E}_{\mathrm{loc}}$ occurs, then the bounds (65), (72) and (73) guarantee

$$\|\widetilde{\boldsymbol{\beta}}_{\mathrm{imd}}^{(t)} - \boldsymbol{\beta}^*\|_{\Sigma} \leq \kappa^{-1} \|\nabla \mathcal{Q}^{(t)}(\boldsymbol{\beta}^*)\|_{\Omega} \leq \kappa^{-1} \big\{ \|\Delta(\widetilde{\boldsymbol{\beta}}^{(t-1)})\|_2 + \|\Delta_1(\widetilde{\boldsymbol{\beta}}^{(t-1)})\|_2 + r_* \big\}, \tag{74}$$

where $\Delta(\cdot)$ and $\Delta_1(\cdot)$ are the random processes defined in (67). For $\widetilde{\boldsymbol{\beta}}^{(t)}$ which minimizes the shifted loss $\widetilde{\mathcal{Q}}^{(t)}(\cdot)$, the first-order condition $\nabla \widetilde{\mathcal{Q}}^{(t)}(\widetilde{\boldsymbol{\beta}}^{(t)}) = \mathbf{0}$ holds, and hence

$$\begin{aligned}
&\|\mathbf{H}(\widetilde{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^*) + \nabla \widehat{\mathcal{Q}}_h(\boldsymbol{\beta}^*)\|_{\Omega} \\
&= \|\Sigma^{-1/2} \{\nabla \widetilde{\mathcal{Q}}^{(t)}(\widetilde{\boldsymbol{\beta}}^{(t)}) - \nabla \widetilde{\mathcal{Q}}^{(t)}(\boldsymbol{\beta}^*) - \mathbf{H}(\widetilde{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^*)\} + \Sigma^{-1/2} \{\nabla \widetilde{\mathcal{Q}}^{(t)}(\boldsymbol{\beta}^*) - \nabla \widehat{\mathcal{Q}}_h(\boldsymbol{\beta}^*)\}\|_2 \\
&\leq \|\Delta_1(\widetilde{\boldsymbol{\beta}}^{(t)})\|_2 + \|\Delta_1(\widetilde{\boldsymbol{\beta}}^{(t-1)})\|_2 + \|\Delta(\widetilde{\boldsymbol{\beta}}^{(t-1)})\|_2.
\end{aligned} \tag{75}$$

In what follows we deal with $\{(\widetilde{\boldsymbol{\beta}}_{\mathrm{imd}}^{(t)}, \widetilde{\boldsymbol{\beta}}^{(t)})\}_{t=1,2,\ldots}$ sequentially, conditioning on $\mathcal{E}_0(r_0) \cap \mathcal{E}_*(r_*) \cap \mathcal{E}_{\mathrm{loc}}$. In view of the basic inequalities (74) and (75), the key is to control the random processes $\Delta(\cdot)$ and $\Delta_1(\cdot)$ as we have done in (68). Define the event

$$\mathcal{F}(r) = \left\{ \sup_{\boldsymbol{\beta} \in \Theta(r)} \big\{ \|\Delta_1(\boldsymbol{\beta})\|_2 + \|\Delta(\boldsymbol{\beta})\|_2 \big\} \leq \delta(x) \cdot r \right\}$$

with $\delta(x) = C\{\sqrt{(p+x)/(nb)} + \sqrt{(p+x)/(Nh)} + b\}$ for some $C > 0$, so that $\mathbb{P}\{\mathcal{F}(r)\} \geq 1 - 2e^{-x}$ for every $0 < r \lesssim b$.

At iteration 1, the bound (74) yields that, conditioned on $\mathcal{E}_0(r_0) \cap \mathcal{E}_*(r_*) \cap \mathcal{E}_{\mathrm{loc}} \cap \mathcal{F}(r_0)$,

$$\|\widetilde{\boldsymbol{\beta}}_{\mathrm{imd}}^{(1)} - \boldsymbol{\beta}^*\|_{\Sigma} \leq r_1 := \kappa^{-1} \delta(x) \cdot r_0 + \kappa^{-1} r_*.$$

The imposed constraints on (b, h, r_0, r_*) ensure that $\kappa^{-1}\delta(x) < 1$, $r_1 < r_{\mathrm{loc}} \asymp b$ and $r_1 \leq r_0$. Via proof by contradiction, we must have $\widetilde{\boldsymbol{\beta}}^{(1)} = \widetilde{\boldsymbol{\beta}}_{\mathrm{imd}}^{(1)} \in \Theta(r_{\mathrm{loc}})$, which in turn certifies

the event $\mathcal{E}_1(r_1) = \{\widetilde{\boldsymbol{\beta}}^{(1)} \in \Theta(r_1)\}$. This, combined with (75) implies that, conditioned on $\mathcal{E}_0(r_0) \cap \mathcal{E}_*(r_*) \cap \mathcal{E}_{\text{loc}} \cap \mathcal{F}(r_0)$,

$$\begin{cases} \|\widetilde{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}^*\|_\Sigma \le \kappa^{-1}\delta(x) \cdot r_0 + \kappa^{-1} r_* = r_1 \le r_0, \\ \|\mathbf{H}(\widetilde{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}^*) + \nabla \widehat{\mathcal{Q}}_h(\boldsymbol{\beta}^*)\|_\Omega \le 2\delta(x) \cdot r_0. \end{cases} \tag{76}$$

Now assume that for some $t \ge 1$, $\widetilde{\boldsymbol{\beta}}^{(t)} \in \Theta(r_t)$ with $r_t := \kappa^{-1}\delta(x) \cdot r_{t-1} + \kappa^{-1} r_* \le r_{t-1}$, and $r_\ell < r_{\text{loc}}$ for all $\ell = 1, \ldots, t$. At iteration $t+1$, applying the general bound (74) again we see that, if event $\mathcal{E}_t(r_t) \cap \mathcal{E}_*(r_*) \cap \mathcal{E}_{\text{loc}} \cap \mathcal{F}(r_t)$ occurs,

$$\|\widetilde{\boldsymbol{\beta}}^{(t+1)}_{\text{imd}} - \boldsymbol{\beta}^*\|_\Sigma \le \kappa^{-1}\delta(x) \cdot r_t + \kappa^{-1} r_*.$$

Set $r_{t+1} = \kappa^{-1}\delta(x) \cdot r_t + \kappa^{-1} r_*$, and note that $r_{t+1} \le \kappa^{-1}\delta(x) \cdot r_{t-1} + \kappa^{-1} r_* = r_t < r_{\text{loc}}$. This means that $\widetilde{\boldsymbol{\beta}}^{(t+1)}_{\text{imd}}$ falls into the interior of $\Theta(r_{\text{loc}})$, which in turn implies $\widetilde{\boldsymbol{\beta}}^{(t+1)} = \widetilde{\boldsymbol{\beta}}^{(t+1)}_{\text{imd}} \in \Theta(r_{\text{loc}})$ and certifies event $\mathcal{E}_{t+1}(r_{t+1})$. Conditioned on $\mathcal{E}_t(r_t) \cap \mathcal{E}_*(r_*) \cap \mathcal{E}_{\text{loc}} \cap \mathcal{F}(r_t)$, we combine the consequence that $\widetilde{\boldsymbol{\beta}}^{(t+1)} \in \Theta(r_{t+1}) \subseteq \Theta(r_t)$ with the general bound (75), thereby obtaining

$$\begin{cases} \|\widetilde{\boldsymbol{\beta}}^{(t+1)} - \boldsymbol{\beta}^*\|_\Sigma \le \kappa^{-1}\delta(x) \cdot r_t + \kappa^{-1} r_* = r_{t+1} \le r_t, \\ \|\mathbf{H}(\widetilde{\boldsymbol{\beta}}^{(t+1)} - \boldsymbol{\beta}^*) + \nabla \widehat{\mathcal{Q}}_h(\boldsymbol{\beta}^*)\|_\Omega \le 2\delta(x) \cdot r_t. \end{cases} \tag{77}$$

Repeat the above argument until we obtain $\widetilde{\boldsymbol{\beta}}^{(T)}$ for some $T \ge 1$. For every $1 \le t \le T$, note that conditioned on $\mathcal{E}_{t-1}(r_{t-1}) \cap \mathcal{E}_*(r_*) \cap \mathcal{E}_{\text{loc}} \cap \mathcal{F}(r_{t-1})$, the event $\mathcal{E}_t(r_t)$ must happen. Therefore, conditioned on $\mathcal{E}_0(r_0) \cap \mathcal{E}_*(r_*) \cap \mathcal{E}_{\text{loc}} \cap \{\cap_{t=0}^{T-1} \mathcal{F}(r_{t-1})\}$, $\widetilde{\boldsymbol{\beta}}^{(T)}$ satisfies the bounds

$$\begin{cases} \|\widetilde{\boldsymbol{\beta}}^{(T)} - \boldsymbol{\beta}^*\|_\Sigma \le \kappa^{-1}\delta(x) \cdot r_{T-1} + \kappa^{-1} r_* =: r_T \le r_{T-1}, \\ \|\mathbf{H}(\widetilde{\boldsymbol{\beta}}^{(T)} - \boldsymbol{\beta}^*) + \nabla \widehat{\mathcal{Q}}_h(\boldsymbol{\beta}^*)\|_\Omega \le 2\delta(x) \cdot r_{T-1}. \end{cases} \tag{78}$$

It is easy to see that $r_t = \{\kappa^{-1}\delta(x)\}^t r_0 + \frac{1-\{\kappa^{-1}\delta(x)\}^t}{1-\kappa^{-1}\delta(x)}\kappa^{-1} r_*$ for $t = 1, \ldots, T$. We thus take $T = \lceil \log(r_0/r_*)/\log(\kappa/\delta(x)) \rceil + 1$, the smallest integer such that $\{\kappa^{-1}\delta(x)\}^{T-1} r_0 \le r_*$.

Finally, conditioned on $\mathcal{E}_0(r_0) \cap \mathcal{E}_*(r_*)$, we combine (76)–(78) with (68), (73) and the union bound to conclude that, with probability at least $1 - (2T+1)e^{-x}$,

$$\begin{cases} \|\widetilde{\boldsymbol{\beta}}^{(T)} - \boldsymbol{\beta}^*\|_\Sigma \le \kappa^{-1}\delta(x) \cdot r_* + \frac{1}{\kappa - \delta(x)} r_* \lesssim r_*, \\ \|\mathbf{H}(\widetilde{\boldsymbol{\beta}}^{(T)} - \boldsymbol{\beta}^*) + \nabla \widehat{\mathcal{Q}}_h(\boldsymbol{\beta}^*)\|_\Omega \le 2\delta(x)\big\{r_* + \frac{1}{\kappa - \delta(x)} r_*\big\} \lesssim \delta(x) \cdot r_* \end{cases}$$

Under the constraints $\sqrt{(p+x)/n} \lesssim b \lesssim 1$ and $\sqrt{(p+x)/N} \lesssim h \le b$, we have $\delta(x) \lesssim b^{1/2}$ and hence $\log(\kappa/\delta(x)) \gtrsim \log(1/b)$. This completes the proof of (15).

B.4 Proof of Theorem 4

Let $\widetilde{\boldsymbol{\beta}}^{(0)}$ be the initial estimator given in (16). For $x > 0$, we can apply either Theorem 2.1 in Pan and Zhou (2021) if $\widetilde{\boldsymbol{\beta}}^{(0)}$ is a local standard QR estimator, or Theorem 3.1 in He *et al.* (2021) with a bandwidth $b \asymp \{(p+x)/n\}^{1/3}$ if $\widetilde{\boldsymbol{\beta}}^{(0)}$ is a local conquer—convolution

smoothed quantile regression—estimator. In either case, $\widetilde{\boldsymbol{\beta}}^{(0)}$ satisfies the bound $\|\widetilde{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^*\|_\Sigma \leq r_0 \asymp \sqrt{(p+x)/n}$ with probability at least $1-2e^{-x}$ as long as $n \gtrsim p+x$. For the second event $\mathcal{E}_*(r_*)$ in (10), it follows from Lemma 16 with $r_* \asymp \sqrt{(p+x)/N} + h^2$ that $\mathbb{P}\{\mathcal{E}_*(r_*)\} \geq 1-e^{-x}$. Putting together the pieces, we conclude that the event $\mathcal{E}_0(r_0) \cap \mathcal{E}_*(r_*)$ occurs with probability at least $1-3e^{-x}$.

Set $x = \log(n \log m)$. Given the specified choice of the bandwidths $b, h > 0$, we have

$$r_0 \asymp \sqrt{\frac{p + \log(n\log m)}{n}}, \quad r_* \asymp \sqrt{\frac{p+\log(n\log m)}{N}}$$

and

$$\sqrt{\frac{p+x}{nb}} + \sqrt{\frac{p+x}{Nh}} + b \asymp \left(\frac{p+\log(n\log m)}{n}\right)^{1/3} + \sqrt{\frac{p+\log(n\log m)}{Nh}}.$$

Finally, applying the high-level result in Theorem 2 yields (17) and (18).

B.5 Proof of Theorem 7

To simplify the presentation, we set $q = p + \log(n\log m)$ throughout the proof. For an arbitrary vector $\boldsymbol{a} \in \mathbb{R}^p$, define the partial sums $S_N = N^{-1/2}\sum_{i=1}^N w_i v_i$ and $S_N^0 = S_N - \mathbb{E}S_N$, where $w_i = \overline{K}(-\varepsilon_i/h) - \tau$ and $v_i = (\mathbf{H}^{-1}\boldsymbol{a})^{\mathrm{T}}\boldsymbol{x}_i$. Recall that $|\mathbb{E}(w_i|\boldsymbol{x}_i)| \leq 0.5 l_1\kappa_2 h^2$ and hence $|\mathbb{E}(w_i v_i)| \leq 0.5 l_1\kappa_2 \|\mathbf{H}^{-1}\boldsymbol{a}\|_\Sigma \cdot h^2$. Now we are ready to prove the normal approximation for $\widetilde{\boldsymbol{\beta}}$. To begin with, we have

$$\begin{aligned}
&|N^{1/2}\boldsymbol{a}^{\mathrm{T}}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) + S_N^0| \\
&\leq N^{1/2}\left|\left\langle \Sigma^{1/2}\mathbf{H}^{-1}\boldsymbol{a}, \Sigma^{-1/2}\mathbf{H}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) + \Sigma^{-1/2}\frac{1}{N}\sum_{i=1}^N \{\overline{K}(-\varepsilon_i/h) - \tau\}\boldsymbol{x}_i \right\rangle\right| + |\mathbb{E}S_N| \\
&\leq N^{1/2}\|\mathbf{H}^{-1}\boldsymbol{a}\|_\Sigma \cdot \left\| \mathbf{H}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) + \frac{1}{N}\sum_{i=1}^N \{\overline{K}(-\varepsilon_i/h) - \tau\}\boldsymbol{x}_i \right\|_\Omega + 0.5 l_1\kappa_2 \|\mathbf{H}^{-1}\boldsymbol{a}\|_\Sigma \cdot N^{1/2}h^2.
\end{aligned}$$

By (18), it follows that with probability at least $1 - Cn^{-1}$,

$$|N^{1/2}\boldsymbol{a}^{\mathrm{T}}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) + S_N^0| \leq C_1 \|\mathbf{H}^{-1}\boldsymbol{a}\|_\Sigma \cdot \{q^{5/6}n^{-1/3} + q(Nh)^{-1/2} + N^{1/2}h^2\}. \tag{79}$$

For the centered partial sum S_N^0, applying the Berry-Esseen inequality (see, e.g. Shevtsova (2013)) yields

$$\sup_{x\in\mathbb{R}} \left|\mathbb{P}\{S_N^0 \leq \mathrm{var}(S_N)^{1/2}x\} - \Phi(x)\right| \leq 0.5 N^{-1/2}\mathrm{var}(wv)^{-3/2}\mathbb{E}|wv - \mathbb{E}(wv)|^3, \tag{80}$$

where $w = \overline{K}(-\varepsilon/h) - \tau$ and $v = (\mathbf{H}^{-1}\boldsymbol{a})^{\mathrm{T}}\boldsymbol{x}$. Following the proof of Lemma 18, it can be shown that $\mathbb{E}(w^2|\boldsymbol{x}) \leq \tau(1-\tau) + (1+\tau)l_0\kappa_2 h^2$ and $|\mathbb{E}(w^2|\boldsymbol{x}) - \tau(1-\tau)| \lesssim h$. Consequently, $\mathrm{var}(wv) = \{\tau(1-\tau) + O(h)\}\|\mathbf{H}^{-1}\boldsymbol{a}\|_\Sigma^2$ and $\mathbb{E}|wv|^3 \leq \max(\tau, 1-\tau)\mathbb{E}(w^2|v|^3) \leq \mu_3\{\tau(1-\tau) + O(h^2)\}\|\mathbf{H}^{-1}\boldsymbol{a}\|_\Sigma^3$, where $\mu_3 = \sup_{\boldsymbol{u}\in\mathbb{S}^{p-1}}\mathbb{E}|\boldsymbol{z}^{\mathrm{T}}\boldsymbol{u}|^3$. Substituting these bounds into (80) gives

$$\sup_{x\in\mathbb{R}} \left|\mathbb{P}\{S_N^0 \leq \mathrm{var}(S_N)^{1/2}x\} - \Phi(x)\right| \leq C_2 N^{-1/2}. \tag{81}$$

Write $\sigma_{\tau,h}^2 = \mathbb{E}\{\overline{K}(-\varepsilon/h) - \tau\}^2\langle \mathbf{H}^{-1}\boldsymbol{a}, \boldsymbol{x}\rangle^2$, and note that $|\mathrm{var}(S_N) - \sigma_{\tau,h}^2| = (\mathbb{E}wv)^2 \leq (0.5l_0\kappa_2 h^2)^2 \cdot \|\mathbf{H}^{-1}\boldsymbol{a}\|_\Sigma^2$. Comparing the distribution functions of two Gaussian random variables shows that

$$\sup_{x\in\mathbb{R}} \big|\Phi\big(x/\mathrm{var}(S_N^0)^{1/2}\big) - \Phi\big(x/\sigma_{\tau,h}\big)\big| \leq C_3 h^4. \tag{82}$$

Let $G \sim \mathcal{N}(0,1)$. Applying the bounds (79), (81) and (82), we conclude that for any $x \in \mathbb{R}$ and $\boldsymbol{a} \in \mathbb{R}^p$,

$$\begin{aligned}
&\mathbb{P}\big\{N^{1/2}\boldsymbol{a}^{\mathrm{T}}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \leq x\big\} \\
&\leq \mathbb{P}\Big[S_N^0 \leq x + C_1\|\mathbf{H}^{-1}\boldsymbol{a}\|_\Sigma \cdot \big\{q^{5/6}n^{-1/3} + q(Nh)^{-1/2} + N^{1/2}h^2\big\}\Big] + Cn^{-1} \\
&\leq \mathbb{P}\Big[\mathrm{var}(S_N^0)^{1/2}G \leq x + C_1\|\mathbf{H}^{-1}\boldsymbol{a}\|_\Sigma \cdot \big\{q^{5/6}n^{-1/3} + q(Nh)^{-1/2} + N^{1/2}h^2\big\}\Big] \\
&\quad + Cn^{-1} + C_2N^{-1/2} \\
&\leq \mathbb{P}\Big[\sigma_{\tau,h}G \leq x + C_1\|\mathbf{H}^{-1}\boldsymbol{a}\|_\Sigma \cdot \big\{q^{5/6}n^{-1/3} + q(Nh)^{-1/2} + N^{1/2}h^2\big\}\Big] \\
&\quad + Cn^{-1} + C_2N^{-1/2} + C_3h^4 \\
&\leq \mathbb{P}\big(\sigma_{\tau,h}G \leq x\big) + Cn^{-1} + \frac{C_1\|\mathbf{H}^{-1}\boldsymbol{a}\|_\Sigma}{(2\pi)^{1/2}\sigma_{\tau,h}}\big\{q^{5/6}n^{-1/3} + q(Nh)^{-1/2} + N^{1/2}h^2\big\} \\
&\quad + C_2N^{-1/2} + C_3h^4.
\end{aligned}$$

A similar argument leads to a series of reverse inequalities. The claimed bound then follows by noting that $|\sigma_{\tau,h}^2 - \tau(1-\tau)\|\mathbf{H}^{-1}\boldsymbol{a}\|_\Sigma^2| \lesssim h$.

B.6 Proof of Proposition 8

Without loss of generality, assume $\mathcal{I}_1 = \{1, \ldots, n\}$, and write $\mathbf{H}_1(\boldsymbol{\beta}) = \mathbb{E}\widehat{\mathbf{H}}_1(\boldsymbol{\beta})$. Consider the change of variable $\boldsymbol{\delta} = \Sigma^{1/2}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)$, so that $\boldsymbol{\beta} \in \Theta(r)$ is equivalent to $\boldsymbol{\delta} \in \mathbb{B}^p(r)$. Recall that $\boldsymbol{z}_i = \Sigma^{-1/2}\boldsymbol{x}_i \in \mathbb{R}^p$ are isotropic random vectors. Define

$$\widehat{\mathbf{H}}(\boldsymbol{\delta}) = \frac{1}{n}\sum_{i=1}^n \phi_b(\varepsilon_i - \boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\delta})\boldsymbol{z}_i\boldsymbol{z}_i^{\mathrm{T}} \quad \text{and} \quad \mathbf{H}(\boldsymbol{\delta}) = \mathbb{E}\big\{\widehat{\mathbf{H}}(\boldsymbol{\delta})\big\}, \tag{83}$$

so that $\widehat{\mathbf{H}}(\boldsymbol{\delta}) = \Sigma^{-1/2}\widehat{\mathbf{H}}_1(\boldsymbol{\beta})\Sigma^{-1/2}$ and $\mathbf{H}(\boldsymbol{\delta}) = \Sigma^{-1/2}\mathbf{H}_1(\boldsymbol{\beta})\Sigma^{-1/2}$, where $\phi_b(u) = (1/b)\phi(u/b)$. For any $\epsilon \in (0, r)$, there exists an ϵ-net $\{\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_{d_\epsilon}\}$ with $d_\epsilon \leq (1 + 2r/\epsilon)^p$ satisfying that, for each $\boldsymbol{\delta} \in \mathbb{B}^p(r)$, there exists some $1 \leq j \leq d_\epsilon$ such that $\|\boldsymbol{\delta} - \boldsymbol{\delta}_j\|_2 \leq \epsilon$. Hence,

$$\begin{aligned}
&\|\widehat{\mathbf{H}}(\boldsymbol{\delta}) - \mathbf{H}(\boldsymbol{\delta})\|_2 \\
&\leq \|\widehat{\mathbf{H}}(\boldsymbol{\delta}) - \widehat{\mathbf{H}}(\boldsymbol{\delta}_j)\|_2 + \|\widehat{\mathbf{H}}(\boldsymbol{\delta}_j) - \mathbf{H}(\boldsymbol{\delta}_j)\|_2 + \|\mathbf{H}(\boldsymbol{\delta}_j) - \mathbf{H}(\boldsymbol{\delta})\|_2 \\
&=: I_1(\boldsymbol{\delta}) + I_2(\boldsymbol{\delta}_j) + I_3(\boldsymbol{\delta}).
\end{aligned}$$

Starting with $I_1(\boldsymbol{\delta})$, note that $|\phi_b(u) - \phi_b(v)| \leq \sup_t |\phi'(t)| \cdot b^{-2}|u - v| \leq (2b)^{-2}|u - v|$ for all $u, v \in \mathbb{R}$. It follows that

$$\begin{aligned} I_1(\boldsymbol{\delta}) &\leq \sup_{\boldsymbol{u},\boldsymbol{v}\in\mathbb{S}^{p-1}} \frac{1}{n}\sum_{i=1}^n |\phi_b(\varepsilon_i - \boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\delta}) - \phi_b(\varepsilon_i - \boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\delta}_j)| \cdot |\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{u} \cdot \boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{v}| \\ &\leq (2b)^{-2} \sup_{\boldsymbol{u},\boldsymbol{v}\in\mathbb{S}^{p-1}} \frac{1}{n}\sum_{i=1}^n |\boldsymbol{z}_i^{\mathrm{T}}(\boldsymbol{\delta} - \boldsymbol{\delta}_j) \cdot \boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{u} \cdot \boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{v}| \\ &\leq (2b)^{-2}\epsilon \cdot \max_{1\leq i\leq n} \|\boldsymbol{z}_i\|_2 \cdot \left\|\frac{1}{n}\sum_{i=1}^n \boldsymbol{z}_i\boldsymbol{z}_i^{\mathrm{T}}\right\|_2. \end{aligned} \tag{84}$$

To bound $\max_{1\leq i\leq n}\|\boldsymbol{z}_i\|_2$, using a standard covering argument we have, for any $\epsilon_1 \in (0,1)$, an ϵ_1-net $\mathcal{N}_{\epsilon_1} \subseteq \mathbb{S}^{p-1}$ with $|\mathcal{N}_{\epsilon_1}| \leq (1 + 2/\epsilon_1)^p$ such that $\max_{1\leq i\leq n}\|\boldsymbol{z}_i\|_2 \leq (1-\epsilon_1)^{-1}\max_{1\leq i\leq n}\max_{\boldsymbol{u}\in\mathcal{N}_{\epsilon_1}} \boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{u}$. Given $1 \leq i \leq n$ and $\boldsymbol{u} \in \mathcal{N}_{\epsilon_1}$, recall that $\mathbb{P}(|\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{u}| \geq \upsilon_1 u) \leq 2e^{-u^2/2}$ for any $u \geq 0$. Taking the union bound over i and $\boldsymbol{u}$, and setting $u = \sqrt{2x + 2\log(2n) + 2p\log(1 + 2/\epsilon_1)}$, we obtain that with probability at least $1 - 2n(1 + 2/\epsilon_1)^p e^{-u^2/2} = 1 - e^{-x}$, $\max_{1\leq i\leq n}\|\boldsymbol{z}_i\|_2 \leq (1-\epsilon_1)^{-1}\upsilon_1\sqrt{2x + 2\log(2n) + 2p\log(1 + 2/\epsilon_1)}$. By minimizing this upper bound with respect to $\epsilon_1 \in (0,1)$, we obtain that with probability at least $1 - e^{-x}$,

$$\max_{1\leq i\leq n}\|\boldsymbol{z}_i\|_2 \lesssim (p + \log n + x)^{1/2}.$$

For $\|(1/n)\sum_{i=1}^n \boldsymbol{z}_i\boldsymbol{z}_i^{\mathrm{T}}\|_2$, it follows from the covering argument along with Bernstein's inequality that, with probability at least $1 - e^{-x}/3$,

$$\left\|\frac{1}{n}\sum_{i=1}^n \boldsymbol{z}_i\boldsymbol{z}_i^{\mathrm{T}} - \mathbf{I}_p\right\|_2 \lesssim \sqrt{\frac{p+x}{n}} \bigvee \frac{p+x}{n}.$$

Plugging the above bounds into (84) yields

$$\sup_{\boldsymbol{\delta}\in\mathbb{B}^p(r)} I_1(\boldsymbol{\delta}) \lesssim (p + \log n + x)^{1/2} b^{-2}\epsilon \tag{85}$$

with probability at least $1 - 2e^{-x}$ as long as $n \gtrsim p + x$. For $I_3(\boldsymbol{\delta})$, it can be similarly obtained that

$$I_3(\boldsymbol{\delta}) \leq (2b)^{-2} \sup_{\boldsymbol{u},\boldsymbol{v}\in\mathbb{S}^{p-1}} \mathbb{E}|\boldsymbol{z}^{\mathrm{T}}(\boldsymbol{\delta} - \boldsymbol{\delta}_j) \cdot \boldsymbol{z}^{\mathrm{T}}\boldsymbol{u} \cdot \boldsymbol{z}^{\mathrm{T}}\boldsymbol{v}| \leq \mu_3(2b)^{-2}\epsilon \tag{86}$$

uniformly over all $\boldsymbol{\delta} \in \mathbb{B}^p(r)$.

Turning to $I_2(\boldsymbol{\delta}_j)$, note that $\widehat{\mathbf{H}}(\boldsymbol{\delta}_j) - \mathbf{H}(\boldsymbol{\delta}_j) = (1/n)\sum_{i=1}^n (1 - \mathbb{E})\phi_{ij}\boldsymbol{z}_i\boldsymbol{z}_i^{\mathrm{T}}$, where $\phi_{ij} = \phi_b(\varepsilon_i - \boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\delta}_j)$ satisfy $|\phi_{ij}| \leq (2\pi)^{-1/2}b^{-1}$ and

$$\mathbb{E}\big(\phi_{ij}^2|\boldsymbol{x}_i\big) = \frac{1}{b^2}\int_{-\infty}^{\infty} \phi^2\left(\frac{\langle \boldsymbol{z}_i, \boldsymbol{\delta}\rangle - t}{b}\right) f_{\varepsilon_i|\boldsymbol{x}_i}(t)\,\mathrm{d}t = \frac{1}{b}\int_{-\infty}^{\infty} \phi^2(u) f_{\varepsilon_i|\boldsymbol{x}_i}(\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\delta} - bu)\,\mathrm{d}u \leq \frac{\bar{f}}{2\pi^{1/2}b}$$

almost surely. Given $\epsilon_2 \in (0, 1/2)$, there exits an ϵ_2-net $\mathcal{M}$ of the sphere $\mathbb{S}^{p-1}$ with $|\mathcal{M}| \le (1+2/\epsilon_2)^p$ such that $\|\widehat{\mathbf{H}}(\boldsymbol{\delta}_j) - \mathbf{H}(\boldsymbol{\delta}_j)\|_2 \le (1-2\epsilon_2)^{-1} \max_{\boldsymbol{u}\in\mathcal{M}} |\boldsymbol{u}^{\mathrm{T}}\{\widehat{\mathbf{H}}(\boldsymbol{\delta}_j) - \mathbf{H}(\boldsymbol{\delta}_j)\}\boldsymbol{u}|$. Given $\boldsymbol{u} \in \mathcal{M}$ and $k = 2, 3, \ldots$, we bound the higher order moments of $\phi_{ij}(\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{u})^2$ by

$$\begin{aligned}
\mathbb{E}|\phi_{ij}(\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{u})^2|^k &\le \bar{f}(2\pi^{1/2}b)^{-1} \cdot \{(2\pi)^{-1/2}b^{-1}\}^{k-2} \upsilon_1^{2k} \cdot 2k \int_0^\infty \mathbb{P}\big(|\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{u}| \ge \upsilon_1 u\big) u^{2k-1} \mathrm{d}u \\
&\le \bar{f}(2\pi^{1/2}b)^{-1} \cdot \{(2\pi)^{-1/2}b^{-1}\}^{k-2} \upsilon_1^{2k} \cdot 4k \int_0^\infty u^{2k-1} e^{-u^2/2} \mathrm{d}u \\
&\le \bar{f}(2\pi^{1/2}b)^{-1} \cdot \{(2\pi)^{-1/2}b^{-1}\}^{k-2} \upsilon_1^{2k} \cdot 2^{k+1} k!.
\end{aligned}$$

In particular, $\mathbb{E}\phi_{ij}^2(\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{u})^4 \le 8\pi^{-1/2}\upsilon_1^4 \bar{f} b^{-1}$, and for each $k \ge 3$, $\mathbb{E}|\phi_{ij}(\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{u})^2|^k \le \frac{k!}{2} \cdot 8\pi^{-1/2}\upsilon_1^4 \bar{f} b^{-1} \cdot (\sqrt{2/\pi}\, \upsilon_1^2 b^{-1})^{k-2}$. Applying Bernstein's inequality and the union bound, we find that for any $u \ge 0$,

$$\begin{aligned}
&\|\widehat{\mathbf{H}}(\boldsymbol{\delta}_j) - \mathbf{H}(\boldsymbol{\delta}_j)\|_2 \\
&\le \frac{1}{1-2\epsilon_2} \max_{\boldsymbol{u}\in\mathcal{M}} \bigg| \frac{1}{n}\sum_{i=1}^n (1 - \mathbb{E})\phi_{ij}(\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{u})^2 \bigg| \le \frac{\upsilon_1^2}{1-2\epsilon_2} \bigg(4\pi^{-1/4} \bar{f}^{1/2} \sqrt{\frac{u}{nb}} + \sqrt{\frac{2}{\pi}} \frac{u}{nb} \bigg)
\end{aligned}$$

with probability at least $1 - 2(1 + 2/\epsilon_2)^p e^{-u} = 1 - e^{\log(2) + p\log(1+2/\epsilon_2) - u}$. Setting $\epsilon_2 = 2/(e^3 - 1)$ and $u = \log(2) + 3p + v$, it follows that with probability at least $1 - e^{-v}$,

$$I_2(\boldsymbol{\delta}_j) \lesssim \sqrt{\frac{p+v}{nb}} + \frac{p+v}{nb}.$$

Once again, taking the union bound over $j = 1, \ldots, d_\epsilon$ and setting $v = p\log(1 + 2r/\epsilon) + x$, we obtain that with probability at least $1 - d_\epsilon e^{-v} \ge 1 - e^{-x}$,

$$\max_{1\le j\le N_\epsilon} I_2(\boldsymbol{\delta}_j) \lesssim \sqrt{\frac{p\log(3er/\epsilon) + x}{nb}} + \frac{p\log(3er/\epsilon) + x}{nb}. \tag{87}$$

Combining (85), (86) and (87), and taking $\epsilon = r/n^2 \in (0, r)$ in the beginning of the proof, we conclude that with probability at least $1 - 3e^{-x}$,

$$\sup_{\boldsymbol{\beta}\in\Theta(r)} \|\widehat{\mathbf{H}}_1(\boldsymbol{\beta}) - \mathbf{H}_1(\boldsymbol{\beta})\|_\Omega \lesssim \sqrt{\frac{p\log n + x}{nb}} + \frac{p\log n + x}{nb} + \frac{(p + \log n + x)^{1/2} r}{(nb)^2}$$

as long as $n \gtrsim p + x$. Moreover, note that for every $\boldsymbol{\beta} \in \Theta(r)$,

$$\begin{aligned}
&\|\mathbf{H}_1(\boldsymbol{\beta}) - \mathbf{H}\|_\Omega \\
&= \bigg\| \mathbb{E} \int_0^1 \int_{-\infty}^\infty \phi(u)\big\{ f_{\varepsilon|\boldsymbol{x}}(t\langle \boldsymbol{x}, \boldsymbol{\beta} - \boldsymbol{\beta}^*\rangle - bu) - f_{\varepsilon|\boldsymbol{x}}(0) \big\} \,\mathrm{d}u \,\mathrm{d}t \cdot \boldsymbol{z}\boldsymbol{z}^{\mathrm{T}} \bigg\|_2.
\end{aligned}$$

By the Lipschitz continuity of $f_{\varepsilon|\boldsymbol{x}}(\cdot)$ and $f'_{\varepsilon|\boldsymbol{x}}(\cdot)$, we have $|f_{\varepsilon|\boldsymbol{x}}(t\langle \boldsymbol{x}, \boldsymbol{\beta} - \boldsymbol{\beta}^*\rangle - bu) - f_{\varepsilon|\boldsymbol{x}}(-bu)| \le l_0 \cdot t \cdot |\boldsymbol{x}^{\mathrm{T}}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)|$ and $|f_{\varepsilon|\boldsymbol{x}}(-bu) - f_{\varepsilon|\boldsymbol{x}}(0) + b f'_{\varepsilon|\boldsymbol{x}}(0) \cdot u| \le |\int_0^{-bu} \{f'_{\varepsilon|\boldsymbol{x}}(v) - f'_{\varepsilon|\boldsymbol{x}}(0)\} \,\mathrm{d}v| \le 0.5 l_1 b^2 u^2$. Plugging these into the above inequality yields

$$\begin{aligned}
&\|\mathbf{H}_1(\boldsymbol{\beta}) - \mathbf{H}\|_\Omega \\
&\le 0.5 l_0 \sup_{\boldsymbol{u}\in\mathbb{S}^{p-1}} \mathbb{E}\big\{ |\boldsymbol{x}^{\mathrm{T}}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)| (\boldsymbol{z}^{\mathrm{T}}\boldsymbol{u})^2 \big\} + 0.5 l_1 b^2 \le 0.5\big(l_0 \mu_3 r + l_1 b^2 \big).
\end{aligned}$$

Putting together the pieces proves the claimed bound, provided that $b \gtrsim (p\log n + x)/n$.

B.7 Proof of Theorem 11

Without loss of generality, assume $\mathcal{I}_1 = \{1, \ldots, n\}$. Let $\mathcal{S} = \mathrm{supp}(\boldsymbol{\beta}^*) \subseteq \{1, \ldots, p\}$ be the true active set with cardinality $|\mathcal{S}| \leq s$, and write $\widetilde{\boldsymbol{\delta}} = \widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$ with $\widetilde{\boldsymbol{\beta}} = \widetilde{\boldsymbol{\beta}}^{(1)}$ for simplicity. By the first-order optimality condition, there exits a subgradient $\widetilde{\boldsymbol{\xi}} \in \partial\|\widetilde{\boldsymbol{\beta}}\|_1$ such that $\nabla\widetilde{\mathcal{Q}}(\widetilde{\boldsymbol{\beta}}) + \lambda \cdot \widetilde{\boldsymbol{\xi}} = \mathbf{0}$ and $\widetilde{\boldsymbol{\xi}}^{\mathrm{T}}\widetilde{\boldsymbol{\beta}} = \|\widetilde{\boldsymbol{\beta}}\|_1$. Hence,

$$\big\langle \widetilde{\boldsymbol{\xi}}, \boldsymbol{\beta}^* - \widetilde{\boldsymbol{\beta}} \big\rangle \leq \|\boldsymbol{\beta}^*\|_1 - \|\widetilde{\boldsymbol{\beta}}\|_1 = \|\boldsymbol{\beta}^*_{\mathcal{S}}\|_1 - \|\widetilde{\boldsymbol{\delta}}_{\mathcal{S}^{\mathrm{c}}}\|_1 - \|\widetilde{\boldsymbol{\delta}}_{\mathcal{S}} + \boldsymbol{\beta}^*_{\mathcal{S}}\|_1 \leq \|\widetilde{\boldsymbol{\delta}}_{\mathcal{S}}\|_1 - \|\widetilde{\boldsymbol{\delta}}_{\mathcal{S}^{\mathrm{c}}}\|_1.$$

This, together with the convexity of $\widetilde{\mathcal{Q}}(\cdot)$, implies

$$\begin{aligned} 0 \leq \overline{D}_{\widetilde{\mathcal{Q}}}(\widetilde{\boldsymbol{\beta}}, \boldsymbol{\beta}^*) &= \big\langle \nabla\widetilde{\mathcal{Q}}(\widetilde{\boldsymbol{\beta}}) - \nabla\widetilde{\mathcal{Q}}(\boldsymbol{\beta}^*), \widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \big\rangle \\ &= \lambda\big\langle \widetilde{\boldsymbol{\xi}}, \boldsymbol{\beta}^* - \widetilde{\boldsymbol{\beta}} \big\rangle - \big\langle \nabla\widetilde{\mathcal{Q}}(\boldsymbol{\beta}^*), \widetilde{\boldsymbol{\delta}} \big\rangle \\ &\leq \lambda\big(\|\widetilde{\boldsymbol{\delta}}_{\mathcal{S}}\|_1 - \|\widetilde{\boldsymbol{\delta}}_{\mathcal{S}^{\mathrm{c}}}\|_1\big) - \big\langle \nabla\widetilde{\mathcal{Q}}(\boldsymbol{\beta}^*), \widetilde{\boldsymbol{\delta}} \big\rangle, \end{aligned} \tag{88}$$

where $\nabla\widetilde{\mathcal{Q}}(\boldsymbol{\beta}^*) = \nabla\widehat{\mathcal{Q}}_{1,b}(\boldsymbol{\beta}^*) - \nabla\widehat{\mathcal{Q}}_{1,b}(\widetilde{\boldsymbol{\beta}}^{(0)}) + \nabla\widehat{\mathcal{Q}}_h(\widetilde{\boldsymbol{\beta}}^{(0)})$. Define gradient-based random processes

$$D_1(\boldsymbol{\beta}) = \nabla\widehat{\mathcal{Q}}_{1,b}(\boldsymbol{\beta}) - \nabla\widehat{\mathcal{Q}}_{1,b}(\boldsymbol{\beta}^*), \quad D(\boldsymbol{\beta}) = \nabla\widehat{\mathcal{Q}}_h(\boldsymbol{\beta}) - \nabla\widehat{\mathcal{Q}}_h(\boldsymbol{\beta}^*),$$

and their means $E_1(\boldsymbol{\beta}) = \mathbb{E}D_1(\boldsymbol{\beta})$ and $E(\boldsymbol{\beta}) = \mathbb{E}D(\boldsymbol{\beta})$. Moreover, let $\mathcal{Q}_h(\boldsymbol{\beta}) = \mathbb{E}(\rho_\tau * K_h)(y - \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta})$ be the population smoothed loss function. It is easy to see that $\mathbb{E}\widehat{\mathcal{Q}}_h(\boldsymbol{\beta}) = \mathcal{Q}_h(\boldsymbol{\beta})$ and $\mathbb{E}\widehat{\mathcal{Q}}_{1,b}(\boldsymbol{\beta}) = \mathcal{Q}_b(\boldsymbol{\beta})$. Then, the gradient $\nabla\widetilde{\mathcal{Q}}(\boldsymbol{\beta}^*)$ can be decomposed as

$$\begin{aligned} \big\{D(\boldsymbol{\beta}) - E(\boldsymbol{\beta})\big\}\Big|_{\boldsymbol{\beta}=\widetilde{\boldsymbol{\beta}}^{(0)}} + \big\{E_1(\boldsymbol{\beta}) - D_1(\boldsymbol{\beta})\big\}\Big|_{\boldsymbol{\beta}=\widetilde{\boldsymbol{\beta}}^{(0)}} + \nabla\widehat{\mathcal{Q}}_h(\boldsymbol{\beta}^*) - \nabla\mathcal{Q}_h(\boldsymbol{\beta}^*) \\ + \big\{E(\boldsymbol{\beta}) - E_1(\boldsymbol{\beta})\big\}\Big|_{\boldsymbol{\beta}=\widetilde{\boldsymbol{\beta}}^{(0)}} + \nabla\mathcal{Q}_h(\boldsymbol{\beta}^*). \end{aligned}$$

For $r > 0$, define the suprema of random processes over the local ℓ_1/ℓ_2 region $\Theta(r) \cap \Lambda$

$$\Pi_1(r) = \sup_{\boldsymbol{\beta}\in\Theta(r)\cap\Lambda} \|D_1(\boldsymbol{\beta}) - E_1(\boldsymbol{\beta})\|_\infty, \quad \Pi(r) = \sup_{\boldsymbol{\beta}\in\Theta(r)\cap\Lambda} \|D(\boldsymbol{\beta}) - E(\boldsymbol{\beta})\|_\infty, \tag{89}$$

and the deterministic quantities

$$\omega(r) = \sup_{\boldsymbol{\beta}\in\Theta(r)} \|E(\boldsymbol{\beta}) - E_1(\boldsymbol{\beta})\|_\Omega, \quad \omega^* = \|\nabla\mathcal{Q}_h(\boldsymbol{\beta}^*)\|_\Omega. \tag{90}$$

If event $\mathcal{E}_*(\lambda_*) \cap \mathcal{E}_0(r_0)$ occurs, then using Hölder's inequality gives

$$|\langle \nabla\widetilde{\mathcal{Q}}(\boldsymbol{\beta}^*), \widetilde{\boldsymbol{\delta}} \rangle| \leq \big\{\Pi(r_0) + \Pi_1(r_0) + \lambda_*\big\} \cdot \|\widetilde{\boldsymbol{\delta}}\|_1 + \{\omega(r_0) + \omega^*\} \cdot \|\widetilde{\boldsymbol{\delta}}\|_\Sigma.$$

Let $\lambda = 2.5(\lambda^* + \varrho)$ with ϱ satisfying

$$\varrho \geq \max\left\{\Pi(r_0) + \Pi_1(r_0), \frac{\omega(r_0) + \omega^*}{s^{1/2}}\right\}, \tag{91}$$

so that $\Pi(r_0) + \Pi_1(r_0) + \lambda_* \leq 0.4\lambda$ and $\omega(r_0) + \omega^* \leq 0.4s^{1/2}\lambda$. Substituting the above bounds into (88) yields $0 \leq 1.4\|\widetilde{\boldsymbol{\delta}}_{\mathcal{S}}\|_1 - 0.6\|\widetilde{\boldsymbol{\delta}}_{\mathcal{S}^{\mathrm{c}}}\|_1 + 0.4s^{1/2}\|\widetilde{\boldsymbol{\delta}}\|_\Sigma$. Consequently, $\|\widetilde{\boldsymbol{\delta}}\|_1 \leq (10/3)\|\widetilde{\boldsymbol{\delta}}_{\mathcal{S}}\|_1 + (2/3)s^{1/2}\|\widetilde{\boldsymbol{\delta}}\|_\Sigma \leq 4s^{1/2}\|\widetilde{\boldsymbol{\delta}}\|_\Sigma$, showing that $\{\widetilde{\boldsymbol{\beta}} \in \Lambda\}$ occurs.

Throughout the rest of the proof, we assume event $\mathcal{E}_*(\lambda_*)\cap\mathcal{E}_0(r_0)$ occurs. Turning to the left-hand side of (88), for $r_{\mathrm{loc}} := b/(4\gamma_{0.25})$, we define $\widetilde{\boldsymbol{\beta}}_\eta = \boldsymbol{\beta}^* + \eta(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)$ with $0 < \eta \le 1$ the same way as in the first paragraph in the proof of Theorem 1. Under the requirement (91) on ϱ, we have $\widetilde{\boldsymbol{\beta}}_\eta \in \Theta(r_{\mathrm{loc}}) \cap \Lambda$ and hence by (88),

$$\overline{D}_{\widetilde{\mathcal{Q}}}(\widetilde{\boldsymbol{\beta}}_\eta, \boldsymbol{\beta}^*) \le \eta \cdot \overline{D}_{\widetilde{\mathcal{Q}}}(\widetilde{\boldsymbol{\beta}}, \boldsymbol{\beta}^*) \le \eta \cdot \big(1.4\lambda\|\widetilde{\boldsymbol{\delta}}_{\mathcal{S}}\|_1 + 0.4s^{1/2}\lambda\|\widetilde{\boldsymbol{\delta}}\|_\Sigma\big) \le 1.8s^{1/2}\lambda \cdot \|\widetilde{\boldsymbol{\beta}}_\eta - \boldsymbol{\beta}^*\|_\Sigma.$$

For the lower bound, Lemma 17 implies

$$\overline{D}_{\widetilde{\mathcal{Q}}}(\widetilde{\boldsymbol{\beta}}_\eta, \boldsymbol{\beta}^*) \ge 0.5\underline{f}\kappa_l \cdot \|\widetilde{\boldsymbol{\beta}}_\eta - \boldsymbol{\beta}^*\|_\Sigma^2$$

with probability at least $1 - e^{-x}$ as long as $(s\log p + x)/n \lesssim b \lesssim 1$. We thus conclude that

$$\|\widetilde{\boldsymbol{\beta}}_\eta - \boldsymbol{\beta}^*\|_\Sigma \le 3.6(\underline{f}\kappa_l)^{-1}s^{1/2}\lambda. \tag{92}$$

It remains to choose a sufficiently large λ, or equivalently ϱ, so that (91) is satisfied. The following two lemmas provide upper bounds on the suprema $\Pi(r_0)$, $\Pi_1(r_0)$ and $\omega(r_0)$, defined in (89) and (90).

Lemma 19 *Assume Conditions (C1), (C2) and (C4) hold. For any $r > 0$ and $x > 0$,*

$$\begin{aligned}\Pi(r) &= \sup_{\boldsymbol{\beta}\in\Theta(r)\cap\Lambda} \|D(\boldsymbol{\beta}) - E(\boldsymbol{\beta})\|_\infty \\ &\le \Big[c_1h^{-1}\sqrt{2s\log(2p)/N} + c_2\sqrt{\{\log(2p)+x\}/(Nh)} + c_3s^{1/2}\{\log(2p)+x\}/(Nh)\Big]\cdot r \end{aligned} \tag{93}$$

with probability at least $1 - e^{-x}$, where $c_1 = 20\kappa_u B^2$, $c_2 = (2\kappa_u\bar{f}\mu_4)^{1/2}\sigma_u$ and $c_3 = (16 + 4/3)\kappa_u B^2$. The same high probability bound, with (N, h) replaced by (n, b), holds for $\Pi_1(r)$.

Lemma 20 *Conditions (C1), (C2) and (C4) ensure that $\omega(r) \le l_0\kappa_1|b-h|r$ for any $r > 0$ and $\omega^* \le l_0\kappa_2h^2/2$, where $\kappa_1 = \int_{-\infty}^{\infty}|u|K(u)\,\mathrm{d}u$ and $\kappa_2 = \int_{-\infty}^{\infty}u^2K(u)\,\mathrm{d}u$.*

Given the bandwidth $0 < h \le b \lesssim 1$, applying Lemmas 19 and 20 yields that

$$\Pi(r_0) + \Pi_1(r_0) \lesssim s^{1/2}r_0 \cdot \left(\frac{1}{b}\sqrt{\frac{\log p + x}{n}} + \frac{1}{h}\sqrt{\frac{\log p + x}{N}}\right)$$

with probability at least $1-2e^{-x}$, and $\omega(r_0)+\omega^* \le l_0(\kappa_1br_0+\kappa_2h^2/2)$. Hence, a sufficiently large ϱ, which is of order

$$\varrho \asymp \max\left[\left\{b^{-1}\sqrt{s\cdot(\log p + x)/n} + h^{-1}\sqrt{s\cdot(\log p + x)/N}\right\}r_0,\ s^{-1/2}\big(br_0 + h^2\big)\right],$$

guarantees that (91) holds with high probability. Consequently, conditioning on $\mathcal{E}_*(\lambda_*) \cap \mathcal{E}_0(r_0)$, the intermediate "estimator" $\widetilde{\boldsymbol{\beta}}_\eta$ satisfies the error bound (92) with probability at least $1 - 3e^{-x}$. We then set $\delta = 3e^{-x} \in (0, 1)$, so that $\log p + x = \log(3p/\delta) \asymp \log(p/\delta)$ and

$$\varrho \asymp \max\left[\left\{b^{-1}\sqrt{s\log(p/\delta)/n} + h^{-1}\sqrt{s\log(p/\delta)/N}\right\}r_0,\ s^{-1/2}\big(br_0 + h^2\big)\right].$$

With the above choice of ϱ, let $b > 14.4\gamma_{0.25}(\underline{f}\kappa_l)^{-1}s^{1/2}\lambda$ so that the right-hand side of (92) is strictly less than r_{loc}. Via proof by contradiction, we must have $\widetilde{\boldsymbol{\beta}} = \widetilde{\boldsymbol{\beta}}_\eta \in \Theta(r_{\mathrm{loc}})$ and hence the bound (92) also applies to $\widetilde{\boldsymbol{\beta}}$, as claimed.

B.8 Proof of Theorem 12

The whole proof will be carried out conditioning on $\mathcal{E}_*(\lambda_*) \cap \mathcal{E}_0(r_0)$ for some prespecified $r_0, \lambda_* > 0$, and write $r_* = s^{1/2}\lambda_*$. Examine the proof of Theorem 11, we see that to obtain the desired error bound for the first iterate $\widetilde{\boldsymbol{\beta}}^{(1)}$, the regularization parameter λ_1 needs to be sufficiently large. Given $\delta \in (0,1)$, we set $\lambda_1 = 2.5(\lambda_* + \varrho_1)$, where $\varrho_1 > 0$ is of order

$$\varrho_1 \asymp \max\left[\left\{b^{-1}\sqrt{s\log(p/\delta)/n} + h^{-1}\sqrt{s\log(p/\delta)/N}\right\} r_0,\ s^{-1/2}(br_0 + h^2)\right]$$

Provided that the bandwidths $b \geq h > 0$ satisfy

$$r_* + \max\left[\left\{b^{-1}s\sqrt{\log(p/\delta)/n} + h^{-1}s\sqrt{\log(p/\delta)/N}\right\} r_0,\ br_0 + h^2\right] \lesssim b \lesssim 1,$$

the first iterate $\widetilde{\boldsymbol{\beta}}^{(1)}$ satisfies $\widetilde{\boldsymbol{\beta}}^{(1)} \in \Lambda$ and

$$\|\widetilde{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}^*\|_\Sigma \leq \underbrace{C_1\left\{b^{-1}s\sqrt{\log(p/\delta)/n} + b + h^{-1}s\sqrt{\log(p/\delta)/N}\right\}}_{=:\,\gamma} \cdot r_0 + C_2(r_* + h^2) =: r_1 \tag{94}$$

with probability at least $1-\delta$, where $\gamma = \gamma(s,p,n,N,h,b,\delta) > 0$ is a contraction factor. With the stated choice of bandwidths $b \asymp s^{1/2}\{\log(p/\delta)/n\}^{1/4}$ and $h \asymp \{s\log(p/\delta)/N\}^{1/4}$, we have

$$\begin{aligned}\gamma &\asymp \left\{s^2\log(p/\delta)/n\right\}^{1/4} + \left\{s^3\log(p/\delta)/N\right\}^{1/4} \\ \text{and}\quad \varrho_1 &\asymp \max\left\{\gamma s^{-1/2} r_0,\ \sqrt{\log(p/\delta)/N}\right\}.\end{aligned}$$

A sufficiently accurate initial estimator—say, $r_0 \lesssim \min\{1, (m/s)^{1/4}\}$—ensures that $s^{1/2}\varrho_1 \lesssim b$. Moreover, we need the local and total sample sizes to be sufficiently large—namely, $n \gtrsim s^2\log(p/\delta)$ and $N \gtrsim s^3\log(p/\delta)$—so that the contraction factor γ is strictly less than 1. As a result, the one-step procedure reduces the estimation error of $\widetilde{\boldsymbol{\beta}}^{(0)}$ by a factor of γ.

For $t = 2, 3, \ldots, T$, define the events $\mathcal{E}_t(r_t) := \{\widetilde{\boldsymbol{\beta}}^{(t)} \in \Theta(r_t) \cap \Lambda\}$ and

$$r_t := \gamma r_{t-1} + C_2(r_* + h^2) = \gamma^t r_0 + C_2\frac{1-\gamma^t}{1-\gamma}(r_* + h^2).$$

At iteration $t \geq 2$, we set $\lambda_t = 3(\lambda_* + \varrho_t)$ with

$$\varrho_t \asymp \max\left\{\gamma s^{-1/2} r_{t-1},\ \sqrt{\log(p/\delta)/N}\right\}.$$

Together, the last two displays imply

$$\varrho_t \asymp \max\left\{\gamma^t s^{-1/2} r_0 + \gamma s^{-1/2}(\lambda_* + h^2)\mathbb{1}(t \geq 2),\ \sqrt{\log(p/\delta)/N}\right\}.$$

Under the stated conditions on r_0 and (n, N), we have $s^{1/2}\varrho_t \lesssim b$ for every $t \geq 2$. Applying Theorem 11 repeatedly, we obtain that conditioned on the event $\mathcal{E}_*(\lambda_*) \cap \mathcal{E}_{t-1}(r_{t-1})$, the t^{th} iterate $\widetilde{\boldsymbol{\beta}}^{(t)}$ satisfies $\widetilde{\boldsymbol{\beta}}^{(t)} \in \Lambda$ and

$$\|\widetilde{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^*\|_\Sigma \leq \gamma r_{t-1} + C_2(r_* + h^2) = r_t = \gamma^t r_0 + C_2\frac{1-\gamma^t}{1-\gamma}(r_* + h^2) \tag{95}$$

with probability at least $1-\delta$.

Note that $r_* = s^{1/2}\lambda^*$ corresponds to the optimal rate under ℓ_2-norm. We thus choose the number of iterations T to be the smallest integer such that $\gamma^T r_0 \le r_*$, that is, $T = \lceil \log(r_0/r_*)/\log(1/\gamma) \rceil$. Applying the union bound over $t = 1, 2, \ldots, T$ yields that conditioned on $\mathcal{E}_*(\lambda_*) \cap \mathcal{E}_0(r_0)$, the T^{th} iterate $\widetilde{\boldsymbol{\beta}}^{(T)}$ satisfies the error bounds

$$\|\widetilde{\boldsymbol{\beta}}^{(T)} - \boldsymbol{\beta}^*\|_\Sigma \lesssim s^{1/2}\lambda^* + h^2 \quad \text{and} \quad \|\widetilde{\boldsymbol{\beta}}^{(T)} - \boldsymbol{\beta}^*\|_1 \lesssim s\lambda^* + s^{1/2}h^2$$

with probability at least $1 - T\delta$. This completes the proof of the theorem.

B.9 Proof of Proposition 13

The proof is similar in spirit to that of Theorem 11, while certain modifications are required. We thus provide a sketch proof for completeness. Note that the shifted loss $\widetilde{\mathcal{Q}}(\cdot)$ in the proof of Theorem 11 shares the Hessian as well as symmetrized Bregman divergence with the local loss $\widehat{\mathcal{Q}}_{1,b}(\cdot)$. With slight abuse of notation, let $\widetilde{\boldsymbol{\delta}} = \widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$ with $\widetilde{\boldsymbol{\beta}} = \widetilde{\boldsymbol{\beta}}^{(0)}$. Inequality (88) implies

$$0 \le \overline{D}_{\widehat{\mathcal{Q}}_{1,b}}(\widetilde{\boldsymbol{\beta}}, \boldsymbol{\beta}^*) \le \lambda_0\big(\|\widetilde{\boldsymbol{\delta}}_{\mathcal{S}}\|_1 - \|\widetilde{\boldsymbol{\delta}}_{\mathcal{S}^{\mathrm{c}}}\|_1\big) - \big\langle \nabla\widehat{\mathcal{Q}}_{1,b}(\boldsymbol{\beta}^*), \widetilde{\boldsymbol{\delta}}\,\big\rangle. \tag{96}$$

By Hölder's inequality,

$$|\langle \nabla\widehat{\mathcal{Q}}_{1,b}(\boldsymbol{\beta}^*), \widetilde{\boldsymbol{\delta}}\,\rangle| \le \|\nabla\widehat{\mathcal{Q}}_{1,b}(\boldsymbol{\beta}^*) - \nabla\mathcal{Q}_{1,b}(\boldsymbol{\beta}^*)\|_\infty \cdot \|\widetilde{\boldsymbol{\delta}}\|_1 + \|\nabla\mathcal{Q}_{1,b}(\boldsymbol{\beta}^*)\|_\Omega \cdot \|\widetilde{\boldsymbol{\delta}}\|_\Sigma,$$

where $\mathcal{Q}_{1,b}(\boldsymbol{\beta}) = \mathbb{E}\widehat{\mathcal{Q}}_{1,b}(\boldsymbol{\beta})$ is the population loss. By the Lipschitz continuity of $f_{\varepsilon|\boldsymbol{x}}(\cdot)$, it can be shown that $\|\nabla\mathcal{Q}_{1,b}(\boldsymbol{\beta}^*)\|_\Omega \le 0.5 l_0\kappa_2 b^2$. Moreover, let the regularization parameter λ_0 satisfy

$$\lambda_0 \ge 2.5\|\nabla\widehat{\mathcal{Q}}_{1,b}(\boldsymbol{\beta}^*) - \nabla\mathcal{Q}_{1,b}(\boldsymbol{\beta}^*)\|_\infty. \tag{97}$$

Then, $|\langle \nabla\widehat{\mathcal{Q}}_{1,b}(\boldsymbol{\beta}^*), \widetilde{\boldsymbol{\delta}}\rangle| \le 0.4\lambda_0\|\widetilde{\boldsymbol{\delta}}\|_1 + 0.5 l_0\kappa_2 b^2\|\widetilde{\boldsymbol{\delta}}\|_\Sigma$. Substituting these into (96) yields

$$\begin{aligned} \overline{D}_{\widehat{\mathcal{Q}}_{1,b}}(\widetilde{\boldsymbol{\beta}}, \boldsymbol{\beta}^*) &\le \big(1.4 s^{1/2}\lambda_0 + 0.5 l_0\kappa_2 b^2\big) \cdot \|\widetilde{\boldsymbol{\delta}}\|_\Sigma \\ \text{and} \quad 0 &\le \lambda_0\big(\|\widetilde{\boldsymbol{\delta}}_{\mathcal{S}}\|_1 - \|\widetilde{\boldsymbol{\delta}}_{\mathcal{S}^{\mathrm{c}}}\|_1\big) + 0.4\lambda_0\|\widetilde{\boldsymbol{\delta}}\|_1 + 0.5 l_0\kappa_2 b^2\|\widetilde{\boldsymbol{\delta}}\|_\Sigma. \end{aligned}$$

The latter implies

$$\|\widetilde{\boldsymbol{\delta}}\|_1 \le (10/3)\|\widetilde{\boldsymbol{\delta}}_{\mathcal{S}}\|_1 + (5/6) l_0\kappa_2\lambda_0^{-1} b^2\|\widetilde{\boldsymbol{\delta}}\|_\Sigma \le L\|\widetilde{\boldsymbol{\delta}}\|_\Sigma \ \text{ with } \ L := (10/3)s^{1/2} + (5/6) l_0\kappa_2\lambda_0^{-1} b^2.$$

Starting from here, we introduce an intermediate "estimator" $\widetilde{\boldsymbol{\beta}}_\eta = \boldsymbol{\beta}^* + \eta(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)$, for some $0 < \eta \le 1$, the same way as in the proof of Theorem 11, so that $\widetilde{\boldsymbol{\beta}}_\eta \in \Theta(r_{\text{loc}})$ with $r_{\text{loc}} = b/(4\gamma_{0.25})$. Since $\widetilde{\boldsymbol{\beta}}_\eta - \boldsymbol{\beta}^* = \eta(\widetilde{\boldsymbol{\beta}}_\eta - \boldsymbol{\beta}^*)$, we also have $\|\widetilde{\boldsymbol{\beta}}_\eta - \boldsymbol{\beta}^*\|_1 \le L\|\widetilde{\boldsymbol{\beta}}_\eta - \boldsymbol{\beta}^*\|_\Sigma$ for the same $L > 1$ given above. Applying Lemma 4.1 in Tan, Wang and Zhou (2022) and Lemma 17, we obtain that with probability at least $1 - e^{-x}$,

$$\overline{D}_{\widehat{\mathcal{Q}}_{1,b}}(\widetilde{\boldsymbol{\beta}}_\eta, \boldsymbol{\beta}^*) \ge 0.5\underline{f}\kappa_l \cdot \|\widetilde{\boldsymbol{\beta}}_\eta - \boldsymbol{\beta}^*\|_\Sigma^2, \ \text{ provided that } n \gtrsim b^{-1}\{L^2\log(p) + x\}. \tag{98}$$

Consequently,

$$0.5\underline{f}\kappa_l \cdot \|\widetilde{\boldsymbol{\beta}}_\eta - \boldsymbol{\beta}^*\|_\Sigma^2 \le \overline{D}_{\widehat{\mathcal{Q}}_{1,b}}(\widetilde{\boldsymbol{\beta}}_\eta, \boldsymbol{\beta}^*) \le \eta \overline{D}_{\widehat{\mathcal{Q}}_{1,b}}(\widetilde{\boldsymbol{\beta}}, \boldsymbol{\beta}^*) \\ \le \big(1.4 s^{1/2}\lambda_0 + 0.5 l_0 \kappa_2 b^2\big) \cdot \|\widetilde{\boldsymbol{\beta}}_\eta - \boldsymbol{\beta}^*\|_\Sigma$$

with probability at least $1 - e^{-x}$. Canceling $\|\widetilde{\boldsymbol{\beta}}_\eta - \boldsymbol{\beta}^*\|_\Sigma$ on both sides yields $\|\widetilde{\boldsymbol{\beta}}_\eta - \boldsymbol{\beta}^*\|_\Sigma \le (\underline{f}\kappa_l)^{-1}(2.8 s^{1/2}\lambda_0 + l_0\kappa_2 b^2)$. Provided that

$$b > 4\gamma_{0.25}(\underline{f}\kappa_l)^{-1}\big(2.8 s^{1/2}\lambda_0 + l_0\kappa_2 b^2\big), \tag{99}$$

$\widetilde{\boldsymbol{\beta}}_\eta$ falls in the interior of $\Theta(r_{\mathrm{loc}})$ (with high probability). Thus we must have $\widetilde{\boldsymbol{\beta}}_\eta = \widetilde{\boldsymbol{\beta}}$, and the same error bound holds for $\widetilde{\boldsymbol{\beta}}$, that is, $\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_\Sigma \lesssim s^{1/2}\lambda_0 + b^2$.

It remains to tune the regularization parameter λ_0 and bandwidth b so that (97) and (99) hold, and to determine the scaling of the sample size required to ensure the lower bound (98). Applying a local version of Lemma 18 yields that with probability at least $1 - e^{-x}$,

$$\|\nabla \widehat{\mathcal{Q}}_{1,b}(\boldsymbol{\beta}^*) - \nabla \mathcal{Q}_{1,b}(\boldsymbol{\beta}^*)\|_\infty \le C(\tau, b)\sqrt{\frac{\log(2p) + x}{n}} + B\max(\tau, 1-\tau)\frac{\log(2p) + x}{3n} \tag{100}$$

for the same constants therein. Given $\delta \in (0,1)$, we take $x = \log(2/\delta)$ in (98) and (100), and set

$$\lambda_0 \asymp \sqrt{\tau(1-\tau)\log(p/\delta)/n},$$

so that (97) holds with high probability. Furthermore, as long as the bandwidth b and sample size n are such that $\sqrt{s\log(p/\delta)/n} \lesssim b \lesssim 1$, both requirements in (98) and (99) are satisfied. This completes the proof of (48).

If in addition $\lambda_0 \ge 1.25 l_0 \kappa_2 s^{-1/2} b^2$, then $\|\widetilde{\boldsymbol{\delta}}\|_1 \le 4 s^{1/2}\|\widetilde{\boldsymbol{\delta}}\|_\Sigma$ and hence $\widetilde{\boldsymbol{\beta}} \in \Lambda$.

Appendix C. Proof of Auxiliary Lemmas

C.1 Proof of Lemma 19

For $r_1, r_2 > 0$, define the parameter set $\Theta_0(r_1, r_2) = \{\boldsymbol{\delta} \in \mathbb{R}^p : \|\boldsymbol{\delta}\|_1 \le r_1, \|\boldsymbol{\delta}\|_\Sigma \le r_2\}$. Consider the change of variable $\boldsymbol{v} = \boldsymbol{\beta} - \boldsymbol{\beta}^*$, so that $\boldsymbol{v} \in \Theta_0(4s^{1/2}r, r)$ for $\boldsymbol{\beta} \in \Theta(r) \cap \Lambda$. Consequently,

$$\sup_{\boldsymbol{\beta}\in\Theta(r_1,r_2)} \|D(\boldsymbol{\beta}) - E(\boldsymbol{\beta})\|_\infty \\ = \max_{1\le j\le p} \sup_{\boldsymbol{v}\in\Theta_0(r_1,r_2)} \left| \frac{1}{N}\sum_{i=1}^N (1-\mathbb{E}) \underbrace{\left\{ \overline{K}\left(\frac{\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{v} - \varepsilon_i}{h}\right) - \overline{K}\left(\frac{-\varepsilon_i}{h}\right)\right\} x_{ij}}_{=:\psi_{ij}(\boldsymbol{v})} \right| =: \max_{1\le j\le p} \Psi_j, \tag{101}$$

where $\Psi_j = \sup_{\boldsymbol{v}\in\Theta_0(r_1,r_2)} |(1/N)\sum_{i=1}^N (1-\mathbb{E})\psi_{ij}(\boldsymbol{v})|$. Since $K(\cdot) = \overline{K}'(\cdot)$ is uniformly bounded,

$$\sup_{\boldsymbol{v}\in\Theta_0(r_1,r_2)} |\psi_{ij}(\boldsymbol{v})| \le \kappa_u B^2 \frac{r_1}{h}.$$

By Bousquet's version of Talagrand's inequality (Bousquet, 2003), we obtain that for any $z > 0$,

$$\Psi_j \le \frac{5}{4}\mathbb{E}\Psi_j + \sup_{\boldsymbol{v}\in\Theta_0(r_1,r_2)} \big\{\mathbb{E}\psi_{ij}^2(\boldsymbol{v})\big\}^{1/2}\sqrt{\frac{2z}{N}} + (4+1/3)\kappa_u B^2 \frac{r_1 z}{Nh} \tag{102}$$

with probability at least $1-2e^{-z}$. For $\boldsymbol{v}\in\Theta_0(r_1,r_2)$,

$$\begin{aligned}
\mathbb{E}\psi_{ij}^2(\boldsymbol{v}) &= \mathbb{E}\left[x_{ij}^2 \int_{-\infty}^{\infty} \big\{\overline{K}\big(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{v}/h - u/h\big) - \overline{K}(-u/h)\big\}^2 f_{\varepsilon|\boldsymbol{x}}(u)\,\mathrm{d}u\right] \\
&= h\,\mathbb{E}\left(x_{ij}^2 \int_{-\infty}^{\infty} \big\{\overline{K}\big(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{v}/h + v\big) - \overline{K}(v)\big\}^2 f_{\varepsilon|\boldsymbol{x}}(-vh)\,\mathrm{d}v\right) \\
&\le \bar{f}h^{-1}\mathbb{E}\left[x_{ij}^2\,(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{v})^2 \int_{-\infty}^{\infty}\left\{\int_0^1 K\big(v + w\boldsymbol{x}^{\mathrm{T}}\boldsymbol{v}/h\big)\,\mathrm{d}w\right\}^2 \mathrm{d}v\right] \\
&\le \bar{f}h^{-1}\mathbb{E}\left(x_{ij}^2\,(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{v})^2 \left[\int_0^1 \left\{\int_{-\infty}^{\infty} K^2\big(v + w\boldsymbol{x}^{\mathrm{T}}\boldsymbol{v}/h\big)\,\mathrm{d}v\right\}^{1/2} \mathrm{d}w\right]^2\right) \\
&\qquad\qquad\qquad\qquad\qquad\qquad \text{(by Minkowski's integral inequality)} \\
&\le \kappa_u\bar{f}\cdot h^{-1}\mathbb{E}\big(x_{ij}\cdot\boldsymbol{x}^{\mathrm{T}}\boldsymbol{v}\big)^2 \le \kappa_u\bar{f}\cdot h^{-1}\big(\mathbb{E}x_{ij}^4\big)^{1/2}\big\{\mathbb{E}(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{v})^4\big\}^{1/2} \le \kappa_u\bar{f}\sigma_{jj}\mu_4\cdot h^{-1}r_2^2,
\end{aligned}$$

where the last inequality uses the bound $\mathbb{E}x_{ij}^4 = \mathbb{E}\langle\Sigma^{-1/2}\boldsymbol{x},\Sigma^{1/2}\boldsymbol{e}_j\rangle^4 \le \mu_4\|\Sigma^{1/2}\boldsymbol{e}_j\|_2^4 = \sigma_{jj}^2\mu_4$.

We next bound the mean $\mathbb{E}\Psi_j$. By Rademacher symmetrization,

$$\mathbb{E}\Psi_j \le 2\mathbb{E}\sup_{\boldsymbol{v}\in\Theta_0(r_1,r_2)}\left|\frac{1}{N}\sum_{i=1}^N e_i\psi_{ij}(\boldsymbol{v})\right| = 2\mathbb{E}\left\{\mathbb{E}_e \sup_{\boldsymbol{v}\in\Theta_0(r_1,r_2)}\left|\frac{1}{N}\sum_{i=1}^N e_i\psi_{ij}(\boldsymbol{v})\right|\right\},$$

where $\mathbb{E}_e$ denotes the conditional expectation over $e_1,\ldots,e_n$ given the remaining variables, and $e_1,\ldots,e_n$ are i.i.d. Rademacher random variables. For each i, write $\psi_{ij}(\boldsymbol{v}) = \varphi_i(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{v})$, where $\varphi_i(\cdot)$ is such that $\varphi_i(0)=0$ and $|\varphi_i(u)-\varphi_i(v)| \le \kappa_u|x_{ij}|\cdot h^{-1}|u-v|$. Then, by Talagrand's contraction principle (see, e.g., Theorem 4.12 in Ledoux and Talagrand (1991)),

$$\begin{aligned}
\mathbb{E}_e \sup_{\boldsymbol{v}\in\Theta_0(r_1,r_2)}\left|\frac{1}{N}\sum_{i=1}^N e_i\psi_{ij}(\boldsymbol{v})\right| &\le 2\kappa_u \max_{1\le i\le N}|x_{ij}|\cdot\mathbb{E}_e \sup_{\boldsymbol{v}\in\Theta_0(r_1,r_2)}\left|\frac{1}{Nh}\sum_{i=1}^N e_i\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{v}\right| \\
&\le 2\kappa_u B\,\frac{r_1}{h}\mathbb{E}_e\left\|\frac{1}{N}\sum_{i=1}^N e_i\boldsymbol{x}_i\right\|_\infty.
\end{aligned}$$

By Hoeffding's moment inequality,

$$\mathbb{E}_e\left\|\frac{1}{N}\sum_{i=1}^N e_i\boldsymbol{x}_i\right\|_\infty \le \max_{1\le k\le p}\left(\frac{1}{N}\sum_{i=1}^N x_{ik}^2\right)^{1/2}\sqrt{\frac{2\log(2p)}{N}}.$$

Putting together the above three inequalities yields

$$\mathbb{E}\Psi_j \le 4\kappa_u B^2 \frac{r_1}{h}\sqrt{\frac{2\log(2p)}{N}} \quad \text{for any } j = 1, \ldots, p. \tag{103}$$

To sum up, we take $z = \log(2p) + x$, $r_1 = 4s^{1/2}r$ and $r_2 = r$ in (102), which combined with (101), (103) and the union bound, completes the proof of (93).

C.2 Proof of Lemma 20

Under the conditional quantile model (1), note that $E(\boldsymbol{\beta}) = \nabla\mathcal{Q}_h(\boldsymbol{\beta}) - \nabla\mathcal{Q}_h(\boldsymbol{\beta}^*)$, where $\mathcal{Q}_h(\boldsymbol{\beta}) = \mathbb{E}\widehat{\mathcal{Q}}_h(\boldsymbol{\beta})$ is the population smoothed loss which is twice-differentiable with Hessian $\nabla^2\mathcal{Q}_h(\boldsymbol{\beta}) = \mathbb{E}\{K_h((\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta} - y)/h)\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\}$. Moreover, define $\mathbf{H}_0 = \mathbb{E}\{f_{\varepsilon|\boldsymbol{x}}(0)\boldsymbol{z}\boldsymbol{z}^{\mathrm{T}}\} = \Sigma^{-1/2}\mathbf{H}\Sigma^{-1/2}$, where $\mathbf{H} = \mathbb{E}\{f_{\varepsilon|\boldsymbol{x}}(0)\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\}$ and $\boldsymbol{z} = \Sigma^{-1/2}\boldsymbol{x}$. By the mean value theorem for vector-valued functions,

$$\begin{aligned}\Sigma^{-1/2}E(\boldsymbol{\beta}) &= \Sigma^{-1/2}\mathbb{E}\int_0^1 \nabla^2\mathcal{Q}_h\big((1-t)\boldsymbol{\beta}^* + t\boldsymbol{\beta}\big)\,\mathrm{d}t\,\Sigma^{-1/2}\cdot\Sigma^{1/2}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)\\ &= \mathbb{E}\int_0^1\int_{-\infty}^{\infty} K(u)f_{\varepsilon|\boldsymbol{x}}(t\cdot\boldsymbol{z}^{\mathrm{T}}\boldsymbol{\delta} - hu)\,\mathrm{d}u\,\mathrm{d}t\cdot\boldsymbol{z}\boldsymbol{z}^{\mathrm{T}}\boldsymbol{\delta},\end{aligned}$$

where $\boldsymbol{\delta} = \Sigma^{1/2}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)$. Similarly,

$$\Sigma^{-1/2}E_1(\boldsymbol{\beta}) = \mathbb{E}\int_0^1\int_{-\infty}^{\infty} K(u)f_{\varepsilon|\boldsymbol{x}}(t\cdot\boldsymbol{z}^{\mathrm{T}}\boldsymbol{\delta} - bu)\,\mathrm{d}u\,\mathrm{d}t\cdot\boldsymbol{z}\boldsymbol{z}^{\mathrm{T}}\boldsymbol{\delta}.$$

where $\mathcal{Q}_{1,b}(\boldsymbol{\beta}) = \mathbb{E}\widehat{\mathcal{Q}}_{1,b}(\boldsymbol{\beta})$. This, together with the Lipschitz continuity of $f_{\varepsilon|\boldsymbol{x}}(\cdot)$ (implied by Condition (C1)), implies that for any $\boldsymbol{\beta} \in \Theta(r)$,

$$\begin{aligned}&\|E(\boldsymbol{\beta}) - E_1(\boldsymbol{\beta})\|_\Omega\\ &\le \sup_{\boldsymbol{u}\in\mathbb{S}^{p-1}}\mathbb{E}\int_0^1\int_{-\infty}^{\infty} K(u)|f_{\varepsilon|\boldsymbol{x}}(t\cdot\boldsymbol{z}^{\mathrm{T}}\boldsymbol{\delta} - hu) - f_{\varepsilon|\boldsymbol{x}}(t\cdot\boldsymbol{z}^{\mathrm{T}}\boldsymbol{\delta} - bu)|\,\mathrm{d}u\,\mathrm{d}t\cdot|\boldsymbol{z}^{\mathrm{T}}\boldsymbol{\delta}\cdot\boldsymbol{z}^{\mathrm{T}}\boldsymbol{u}|\\ &\le l_0\int_{-\infty}^{\infty}|u|K(u)\,\mathrm{d}u\cdot|b-h|\sup_{\boldsymbol{u}\in\mathbb{S}^{p-1}}\big\{\mathbb{E}(\boldsymbol{z}^{\mathrm{T}}\boldsymbol{u})^2\big\}^{1/2}\|\boldsymbol{\delta}\|_\Sigma = l_0\kappa_1|b-h|r,\end{aligned}$$

as claimed.

Turning to $\Sigma^{-1/2}\nabla\mathcal{Q}_h(\boldsymbol{\beta}^*) = \mathbb{E}\{\overline{K}(-\varepsilon/h) - \tau\}\boldsymbol{z}$, by integration by parts we get

$$\begin{aligned}\mathbb{E}\big\{\overline{K}(-\varepsilon/h)|\boldsymbol{x}\big\} &= \int_{-\infty}^{\infty}\overline{K}(-t/h)\,\mathrm{d}F_{\varepsilon|\boldsymbol{x}}(t)\\ &= -\frac{1}{h}\int_{-\infty}^{\infty}K(-t/h)F_{\varepsilon|\boldsymbol{x}}(t)\,\mathrm{d}t = \int_{-\infty}^{\infty}K(u)F_{\varepsilon|\boldsymbol{x}}(-hu)\,\mathrm{d}t\\ &= \tau + \int_{-\infty}^{\infty}K(u)\int_0^{-hu}\big\{f_{\varepsilon|\boldsymbol{x}}(t) - f_{\varepsilon|\boldsymbol{x}}(0)\big\}\,\mathrm{d}t\,\mathrm{d}u.\end{aligned}$$

Combined with the Lipschitz continuity of $f_{\varepsilon|\boldsymbol{x}}(\cdot)$, this implies

$$\|\nabla\mathcal{Q}_h(\boldsymbol{\beta}^*)\|_\Omega = \sup_{\boldsymbol{u}\in\mathbb{S}^{p-1}}\mathbb{E}\int_{-\infty}^{\infty}K(u)\int_0^{-hu}\big\{f_{\varepsilon|\boldsymbol{x}}(t) - f_{\varepsilon|\boldsymbol{x}}(0)\big\}\,\mathrm{d}t\,\mathrm{d}u\cdot\boldsymbol{z}^{\mathrm{T}}\boldsymbol{u} \le \frac{1}{2}l_0\kappa_2 h^2,$$

thus completing the proof.

Appendix D. Estimation at Extreme Quantile Levels

As highlighted in the introduction, the causal mechanisms underpinning extreme behavior are of high relevance in numerous fields, and QR at extreme quantile levels operationalizes attempts to understand these. Rather than treating variables on an equal footing as Engelke and Hitz (2020), QR singles out a particular variable for which understanding is sought. Just as the statistical aspects of extreme value theory are challenged by the limitation of data beyond extreme thresholds, QR coefficients at extreme quantiles are notoriously hard to estimate. The following minor adaptation of our procedure improves its performance at extreme quantile levels.

Recall from Section 2.1 that the conquer method is evolved from a smoothed estimating equation approach (Kaplan and Sun, 2017). The latter constructs a smoothed sample analog of the moment condition. As observed by both Fernandes, Guerre and Horta (2021) and He *et al.* (2021), the smoothing bias primarily affects the intercept estimation especially in the random design setting. Now let us take a closer look at the moment condition. For every p-vector $\boldsymbol{u} = (u_1, \ldots, u_p)^{\mathrm{T}}$, we use $\boldsymbol{u}_- \in \mathbb{R}^{p-1}$ to denote its $(p-1)$-subvector with the first coordinate removed, i.e. $\boldsymbol{u}_- = (u_2, \ldots, u_p)^{\mathrm{T}}$. Then, the first-order moment condition can be written as

$$\begin{cases} \mathbb{E}\{\mathbb{1}(y < \boldsymbol{x}_-^{\mathrm{T}}\boldsymbol{\beta}_- + \beta_1) - \tau\} = 0, \\ \mathbb{E}\{\mathbb{1}(y < \boldsymbol{x}_-^{\mathrm{T}}\boldsymbol{\beta}_- + \beta_1) - \tau\}x_j = 0, \quad j = 2, \ldots, p, \end{cases}$$

whose sample counterpart is

$$\begin{cases} \sum_{i=1}^N \{\mathbb{1}(y_i < \boldsymbol{x}_{i,-}^{\mathrm{T}}\boldsymbol{\beta}_- + \beta_1) - \tau\} = 0, \\ \sum_{i=1}^N \{\mathbb{1}(y_i < \boldsymbol{x}_{i,-}^{\mathrm{T}}\boldsymbol{\beta}_- + \beta_1) - \tau\}x_{ij} = 0, \quad j = 2, \ldots, p. \end{cases} \tag{104}$$

To find the solution of the above system of equations, note that given $\boldsymbol{\beta}_- \in \mathbb{R}^{p-1}$, the first equation can be (approximately) solved by taking β_1 to be the sample τ-quantile of $\{y_i - \boldsymbol{x}_{i,-}^{\mathrm{T}}\boldsymbol{\beta}_-\}_{i=1}^N$, which only allows for an error $1/N$. The main difficulty then arises from solving the remaining $p-1$ equations, for which analytical solutions do not exist. To mitigate the smoothing bias of conquer for intercept estimation, we consider a hybrid estimating equation approach that solves

$$\begin{cases} \sum_{i=1}^N \{\mathbb{1}(y_i < \boldsymbol{x}_{i,-}^{\mathrm{T}}\boldsymbol{\beta}_- + \beta_1) - \tau\} = 0, \\ \sum_{i=1}^N \{\overline{K}(-(y_i - \beta_1 - \boldsymbol{x}_{i,-}^{\mathrm{T}}\boldsymbol{\beta}_-)/h) - \tau\}\boldsymbol{x}_{i,-} = \mathbf{0}_{p-1}, \end{cases} \tag{105}$$

where $\overline{K}(u) = \int_{-\infty}^u K(v)\,\mathrm{d}v$ for some kernel function $K(\cdot)$. Note that, given $\boldsymbol{\beta}_- \in \mathbb{R}^{p-1}$, the first equation in (105) can be solved by taking β_1 to be the sample τ-quantile of $\{y_i - \boldsymbol{x}_{i,-}^{\mathrm{T}}\boldsymbol{\beta}_-\}_{i=1}^N$, and given β_1, solving the second vector equation is equivalent to minimizing the conquer loss $\boldsymbol{\beta}_- \mapsto \widehat{\mathcal{Q}}_h(\beta_1, \boldsymbol{\beta}_-)$. This motivates the following iterative procedure, starting at iteration 0 with initial estimates $\beta_1^{(0)}$ and $\boldsymbol{\beta}_-^{(0)}$ of the intercept and slope coefficients, respectively. The procedure involves two steps. At iteration $t = 1, 2, \ldots$:

Refitted intercept. Using the current slope coefficients estimate $\boldsymbol{\beta}_-^{(t-1)}$, we compute the residuals $r_i^{(t-1)} = y_i - \boldsymbol{x}_{i,-}^{\mathrm{T}}\boldsymbol{\beta}_-^{(t-1)}$, and then update the intercept $\beta_1^{(t)}$ as the sample τ-quantile of $\{r_i^{(t-1)}\}_{i=1}^N$.

Adjusted conquer. With a refitted intercept $\beta_1^{(t)}$, we take any solution to the optimization problem

$$\min_{\boldsymbol{\beta}_- \in \mathbb{R}^{p-1}} \widehat{\mathcal{Q}}_h(\beta_1^{(t)}, \boldsymbol{\beta}_-) = \min_{\boldsymbol{\beta}_- \in \mathbb{R}^{p-1}} \frac{1}{N} \sum_{i=1}^{N} (\rho_\tau * K_h)(y_i - \beta_1^{(t)} - \boldsymbol{x}_{i,-}^{\mathrm{T}} \boldsymbol{\beta}_-)$$

as the updated slope estimator, denoted by $\widehat{\boldsymbol{\beta}}_-^{(t)}$. We refer to the above method as *two-step conquer*.

Using two-step conquer, in Algorithm 4 we present a modified multi-round distributed algorithm, which is particularly suited for extreme quantile regressions with τ close to either 0 or 1.

Appendix E. Additional Simulation Studies

In this section, we provide numerical studies for score-based confidence sets to complement those in Section 4.2 of the paper. Specifically, in each of 200 Monte Carlo replications, $p = 10$ covariates are generated at random from a uniform distribution on $[-1, 1]$ and a response variable is generated according to the linear heteroscedastic model

$$y_i = \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}^* + (0.25 x_{i1} + 0.25 x_{i2} + 0.75)\{\varepsilon_i - F_{\varepsilon_i}^{-1}(\tau)\}, \quad i = 1, \ldots, N,$$

where $\tau = 0.9$ and ε_i is drawn from the t-distribution with 1.5 degrees of freedom. The intercept β_0^* is taken as 2 and all other elements of $\boldsymbol{\beta}^*$ as unity.

For $n = 200$ and $m = 100$, Tables 4 and 5 report the simulated coverage probabilities and mean width for each confidence set construction described in Section 2.3. In addition to the aforementioned methods, the construction in equation (32) is referred to as CE-Score.

	β_1^*	β_2^*	β_3^*	β_4^*	β_5^*	β_6^*	β_7^*	β_8^*	β_9^*	β_{10}^*
DC-Normal	0.810	0.805	0.995	1.000	0.985	0.995	0.995	0.990	1.000	0.995
CE-Normal	0.935	0.865	0.925	0.930	0.895	0.940	0.895	0.925	0.910	0.910
CE-Boot (a)	0.970	0.940	0.960	0.945	0.940	0.975	0.950	0.955	0.950	0.955
CE-Boot (b)	0.965	0.910	0.945	0.950	0.935	0.975	0.935	0.955	0.930	0.950
CE-Score	0.960	0.905	0.960	0.970	0.915	0.980	0.920	0.955	0.945	0.950

Table 4: Monte Carlo coverage probabilities for the case when $p = 10$, $n = 200$, and $m = 100$.

Apart from those based on the averaging estimator $\widehat{\boldsymbol{\beta}}^{\mathrm{dc}}$, which has appreciable under-coverage for the fist two coefficients, all constructions are broadly comparable in terms of coverage probability. Confidence intervals based on the score statistic are considerably narrower than the others, although at higher computational cost due to the need to calculate $\widehat{T}_k(c_k)$ from equation (32) over a grid of c_k values.

Tables 6 and 7 for $n = 200$ and $m = 200$ are qualitatively similar. For roughly similar coverage, the higher total sample size $N = nm$ reduces the widths of the confidence intervals. The larger value of m also has the effect of reducing the DC-Normal coverage probability

Algorithm 4 Efficient Distributed Quantile Regression via Two-Step Conquer.

Input: data batches $\{(y_i, \boldsymbol{x}_i)\}_{i\in\mathcal{I}_j}$, $j = 1, \ldots, m$, stored at m sites, quantile level $\tau \in (0,1)$, bandwidths $b, h > 0$, initialization $\widetilde{\boldsymbol{\beta}}^{(0)} \in \mathbb{R}^p$, maximum number of iterations T, $g_0 = 1$.

1: **for** $t = 1, 2 \ldots, T$ **do**

2: Broadcast $\widetilde{\boldsymbol{\beta}}_{-}^{(t-1)} \in \mathbb{R}^{p-1}$ to all local machines.

3: **for** $j = 1, \ldots, m$ **do**

4: At the jth site, compute the sample τ-quantile of $\{\widehat{r}_i^{(t-1)} := y_i - \langle \boldsymbol{x}_{i,-}, \widetilde{\boldsymbol{\beta}}_{-}^{(t-1)} \rangle\}_{i\in\mathcal{I}_j}$, denoted by $\widehat{q}_j^{(t-1)}$, and send it the master (first) machine.

5: **end for**

6: Calculate $\widehat{q}^{(t)} = (1/m)\sum_{j=1}^m \widehat{q}_j^{(t-1)}$ on the master, and send it to every local machine.

7: **for** $j = 1, \ldots, m$ **do**

8: On the jth machine, compute the gradient vector

$$\widehat{\boldsymbol{g}}_{j,h}^{(t-1)} = -\frac{1}{n}\sum_{i\in\mathcal{I}_j} \ell_h'(\widehat{r}_i^{(t-1)} - \widehat{q}^{(t)})\boldsymbol{x}_{i,-} \in \mathbb{R}^{p-1},$$

and send it to the master.

9: **end for**

10: On the master machine, calculate

$$\widehat{\boldsymbol{g}}_h^{(t-1)} = \frac{1}{m}\sum_{j=1}^m \widehat{\boldsymbol{g}}_{j,h}^{(t-1)} \quad \text{and} \quad g_t = \|\widehat{\boldsymbol{g}}_h^{(t-1)}\|_\infty.$$

11: **if** $g_t > g_{t-1}$ or $g_t < 10^{-5}$ **break**

12: **otherwise** Calculate $\widehat{\boldsymbol{g}}_{1,b}^{(t-1)} = (-1/n)\sum_{i\in\mathcal{I}_1} \ell_b'(\widehat{r}_i^{(t-1)} - \widehat{q}^{(t)})\boldsymbol{x}_{i,-}$, and solve the shifted conquer loss minimization

$$\widehat{\boldsymbol{\theta}}^{(t)} \in \underset{\boldsymbol{\theta}\in\mathbb{R}^{p-1}}{\operatorname{argmin}} \ \widehat{\mathcal{Q}}_{1,b}(\widehat{q}^{(t)}, \boldsymbol{\theta}) - \langle \widehat{\boldsymbol{g}}_{1,b}^{(t-1)} - \widehat{\boldsymbol{g}}_h^{(t-1)}, \boldsymbol{\theta} \rangle$$

on the master machine. Define $\widetilde{\boldsymbol{\beta}}^{(t)} = (\widehat{q}^{(t)}, (\widehat{\boldsymbol{\theta}}^{(t)})^{\mathrm{T}})^{\mathrm{T}}$ as the t^{th} iterate.

13: **end for**

Output: $\widetilde{\boldsymbol{\beta}}^{(T)}$.

due to the bias in $\widehat{\boldsymbol{\beta}}^{\mathrm{dc}}$. This bias does not disappear asymptotically in m but rather the variation of $\widehat{\boldsymbol{\beta}}^{\mathrm{dc}}$ around the wrong point is diminished, leading to poor coverage.

Further simulation results for $n = 400$ with $m = 100$ and $m = 200$ are reported in Tables 8–11. Similar conclusions can be drawn.

	β_1^*	β_2^*	β_3^*	β_4^*	β_5^*	β_6^*	β_7^*	β_8^*	β_9^*	β_{10}^*
DC-Normal	0.359	0.344	0.338	0.336	0.348	0.347	0.344	0.336	0.337	0.340
CE-Normal	0.207	0.197	0.191	0.187	0.192	0.192	0.191	0.186	0.189	0.189
CE-Boot (a)	0.238	0.234	0.216	0.217	0.228	0.225	0.222	0.216	0.216	0.219
CE-Boot (b)	0.222	0.215	0.206	0.203	0.210	0.209	0.208	0.202	0.204	0.204
CE-Score	0.162	0.162	0.162	0.162	0.162	0.161	0.161	0.161	0.161	0.161

Table 5: Monte Carlo mean width of the constructed confidence intervals for the case when $p = 10$, $n = 200$, and $m = 100$.

	β_1^*	β_2^*	β_3^*	β_4^*	β_5^*	β_6^*	β_7^*	β_8^*	β_9^*	β_{10}^*
DC-Normal	0.595	0.615	0.990	0.995	1.000	0.990	0.995	1.000	0.990	0.995
CE-Normal	0.895	0.925	0.935	0.950	0.915	0.910	0.940	0.950	0.920	0.900
CE-Boot (a)	0.955	0.945	0.960	0.985	0.960	0.960	0.985	0.980	0.955	0.945
CE-Boot (b)	0.945	0.935	0.950	0.975	0.955	0.935	0.985	0.975	0.950	0.930
CE-Score	0.955	0.955	0.950	0.970	0.955	0.935	0.970	0.980	0.940	0.965

Table 6: Monte Carlo coverage probabilities for the case when $p = 10$, $n = 200$, and $m = 200$.

	β_1^*	β_2^*	β_3^*	β_4^*	β_5^*	β_6^*	β_7^*	β_8^*	β_9^*	β_{10}^*
DC-Normal	0.259	0.257	0.257	0.247	0.251	0.251	0.254	0.244	0.252	0.245
CE-Normal	0.149	0.145	0.138	0.137	0.135	0.138	0.138	0.134	0.135	0.134
CE-Boot (a)	0.173	0.173	0.165	0.162	0.160	0.164	0.164	0.158	0.162	0.160
CE-Boot (b)	0.165	0.163	0.155	0.154	0.150	0.154	0.154	0.150	0.152	0.151
CE-Score	0.114	0.114	0.113	0.113	0.113	0.113	0.113	0.112	0.113	0.112

Table 7: Monte Carlo mean width of the constructed confidence intervals for the case when $p = 10$, $n = 200$, and $m = 200$.

	β_1^*	β_2^*	β_3^*	β_4^*	β_5^*	β_6^*	β_7^*	β_8^*	β_9^*	β_{10}^*
DC-Normal	0.765	0.715	0.995	1.000	0.995	0.980	0.990	1.000	0.985	1.000
CE-Normal	0.965	0.950	0.935	0.960	0.940	0.920	0.960	0.960	0.960	0.960
CE-Boot (a)	0.990	0.970	0.970	0.980	0.950	0.940	0.965	0.975	0.960	0.955
CE-Boot (b)	0.975	0.965	0.945	0.975	0.960	0.930	0.975	0.965	0.960	0.975
CE-Score	0.950	0.950	0.945	0.970	0.950	0.935	0.980	0.980	0.930	0.965

Table 8: Monte Carlo coverage probabilities ($p = 10$, $n = 400$, $m = 100$).

	β_1^*	β_2^*	β_3^*	β_4^*	β_5^*	β_6^*	β_7^*	β_8^*	β_9^*	β_{10}^*
DC-Normal	0.191	0.189	0.187	0.184	0.184	0.184	0.185	0.180	0.185	0.182
CE-Normal	0.141	0.139	0.132	0.130	0.130	0.130	0.131	0.130	0.133	0.131
CE-Boot (a)	0.153	0.154	0.147	0.144	0.144	0.144	0.146	0.143	0.146	0.145
CE-Boot (b)	0.147	0.147	0.139	0.138	0.137	0.137	0.139	0.137	0.140	0.138
CE-Score	0.114	0.114	0.112	0.113	0.112	0.113	0.113	0.112	0.112	0.113

Table 9: Monte Carlo mean width ($p = 10$, $n = 400$, $m = 100$).

	β_1^*	β_2^*	β_3^*	β_4^*	β_5^*	β_6^*	β_7^*	β_8^*	β_9^*	β_{10}^*
DC-Normal	0.495	0.450	0.985	0.995	0.985	0.990	0.985	0.980	0.980	0.990
CE-Normal	0.945	0.915	0.940	0.960	0.950	0.935	0.925	0.950	0.930	0.930
CE-Boot (a)	0.965	0.940	0.960	0.980	0.975	0.950	0.960	0.980	0.945	0.970
CE-Boot (b)	0.955	0.935	0.955	0.975	0.970	0.950	0.945	0.970	0.930	0.960
CE-Score	0.930	0.930	0.965	0.965	0.955	0.930	0.965	0.950	0.955	0.960

Table 10: Monte Carlo coverage probabilities ($p = 10$, $n = 400$, $m = 200$).

	β_1^*	β_2^*	β_3^*	β_4^*	β_5^*	β_6^*	β_7^*	β_8^*	β_9^*	β_{10}^*
DC-Normal	0.131	0.129	0.128	0.127	0.127	0.126	0.129	0.124	0.124	0.125
CE-Normal	0.097	0.096	0.091	0.091	0.092	0.091	0.092	0.089	0.089	0.090
CE-Boot (a)	0.106	0.105	0.100	0.101	0.101	0.100	0.103	0.099	0.097	0.099
CE-Boot (b)	0.103	0.102	0.097	0.097	0.097	0.097	0.098	0.095	0.094	0.095
CE-Score	0.078	0.078	0.077	0.077	0.077	0.077	0.077	0.078	0.077	0.077

Table 11: Monte Carlo mean width ($p = 10$, $n = 400$, $m = 200$).

References

Barzilai, J. and Borwein, J. M. (1988). Two-point step size gradient methods. *IMA Journal of Numerical Analysis* **8** 141–148.

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** 1352–1382.

Belloni, A. and Chernozhukov, V. (2011). ℓ_1-regularized quantile regression in high-dimensional sparse models. *Annals of Statistics* **39** 82–130.

Belloni, A., Chernozhukov, V., Fernández-Val, I. and Hansen, C. (2017). Program evaluation and causal inference with high-dimensional data. *Econometrica* **85** 233-298.

Bickel, P. J. (1975). One-step Huber estimates in the linear model. *Journal of the American Statistical Association* **70** 428–434.

Bousquet, O. (2003). Concentration inequalities for sub-additive functions using the entropy method. *In Stochastic Inequalities and Applications. Progress in Probability* **56** 213–247. Birkhäuser, Basel.

Boyd, S., Parikh, N., Chu, E., Peleato, B. and Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning* **3** 1–122.

Bradic, J. and Kolar, M. (2017). Uniform inference for high-dimensional quantile regression: Linear functionals and regression rank scores. *arXiv preprint arXiv:1702.06209.*

Chen, X., Liu, W. and Zhang, Y. (2021). First-order Newton-type estimator for distributed estimation and inference. *Journal of American Statistical Association* DOI: 10.1080/01621459.2021.1891925.

Chen, X., Liu, W. and Zhang, Y. (2019). Quantile regression under memory constraint. *Annals of Statistics* **47** 3244–3273.

Cochran, W. G. (1938). The omission or addition of an independent variable in multiple linear regression. *Supplement to the Journal of the Royal Statistical Society* **5** 171–176.

Cox, D. R. (2007). On a generalization of a result of W. G. Cochran. *Biometrika* **94** 755–759.

de la Peña, V. H., Lai, T. L. and Shao, Q.-M. (2009). *Self-Normalized Processes: Theory and Statistical Applications.* Springer, Berlin.

Douglas, J. and Rachford, H. H. (1956). On the numerical solution of heat conduction problems in two and three space variables. *Transactions of the American Mathematical Society* **82** 421–439.

Efron, B. (1969). Student's t-test under symmetry conditions. *Journal of the American Statistical Association* **64** 1278–1302.

Engelke, S. and Hitz, A. S. (2020). Graphical models for extremes. *Journal of the Royal Statistical Society: Series B* **82** 869–931.

Fan, J., Guo, Y. and Wang, K. (2021). Communication-efficient accurate statistical estimation. *Journal of the American Statistical Association* DOI: 10.1080/01621459.2021.1969238.

Fan, J., Liu, H., Sun, Q. and Zhang, T. (2018). I-LAMM for sparse learning: Simultaneous control of algorithmic complexity and statistical error. *Annals of Statistics* **46** 814–841.

Fernandes, M., Guerre, E. and Horta, E. (2021). Smoothing quantile regressions. *Journal of Business and Economics Statistics* **39** 338–357.

Fieller, E. C. (1954). Some problems in interval estimation. *Journal of the Royal Statistical Society: Series B* **16** 175–185.

Gabay, D. and Mercier, B. (1976). A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Computers & Mathematics with Applications* **2** 17–40.

Galvao, A. F. and Kato, K. (2016). Smoothed quantile regression for panel data. *Journal of Econometrics* **193** 92–112.

Gu, Y., Fan, J., Kong, L., Ma, S. and Zou, H. (2018). ADMM for high-dimensional sparse regularized quantile regression. *Technometrics* **60** 319–331.

Hall, P. and Sheather, S. J. (1988). On the distribution of a studentized quantile. *Journal of the Royal Statistical Society: Series B* **50** 381–391.

He, X., Pan, X., Tan, K. M. and Zhou, W.-X. (2021). Smoothed quantile regression with large-scale inference. *Journal of Econometrics* DOI: 10.1016/j.jeconom.2021.07.010.

Horowitz, J. L. (1998). Bootstrap methods for median regression models. *Econometrica* **66** 1327–1351.

Hunter, D. R. and Lange, K. (2000). Quantile regression via an MM algorithm. *Journal of Computational and Graphical Statistics* **9** 60–77.

Jiang, R. and Yu, K. (2021). Smoothing quantile regression for a distributed system. *Neurocomputing* **466** 311–326.

Jordan, M. I., Lee, J. D. and Yang, Y. (2019). Communication-efficient distributed statistical inference. *Journal of the American Statistical Association* **114** 668–681.

Kaplan, D. M. and Sun, Y. (2017). Smoothed estimating equations for instrumental variables quantile regression. *Econometric Theory* **33** 105–157.

Koenker, R. (2005). *Quantile Regression.* Cambridge University Press, Cambridge.

Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrica* **46** 33–50.

Koenker, R., Chernozhukov, V., He, X. and Peng, L. (2017). *Handbook of Quantile Regression.* CRC Press, New York.

Lan, G., Lee, S. and Zhou, Y. (2020). Communication-efficient algorithms for decentralized and stochastic optimization. *Mathematical Programming* **180** 237–284.

Ledoux, M. and Talagrand, M. (1991). *Probability in Banach Spaces: Isoperimetry and Processes*. Springer-Verlag, Berlin.

Lee, J., Sun, Y., Liu, Q. and Taylor, J. (2017). Communication-efficient sparse regression. *Journal of Machine Learning Research* **18**(5): 1–30.

Li, T., Sahu, A. K., Talwalkar, A. and Smith, V. (2020). Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine* **37** 50–60.

Li, Y. and Zhu, J. (2008). L_1-norm quantile regression. *Journal of Computational and Graphical Statistics* **17** 163–185.

Pan, X. and Zhou, W.-X. (2021). Multiplier bootstrap for quantile regression: Nonasymptotic theory under random design. *Information and Inference: A Journal of the IMA* **10** 813–861.

Portnoy, S. and Koenker, R. (1997). The Gaussian hare and the Laplacian tortoise: Computability of squared-error versus absolute-error estimators. *Statistical Science* **12** 279–300.

Powell, J. L. (1991). Estimation of monotonic regression models under quantile restrictions. In *Nonparametric and Semiparametric Methods in Econometrics* (eds W. Barnett, J. Powell and G. Tauchen). Cambridge Univ. Press, Cambridge.

Shamir, O., Srebro, N. and Zhang, T. (2014). Communication efficient distributed optimization using an approximate Newton-type method. In *Proceedings of the 31st International Conference on Machine Learning* **32** 1000–1008.

Shevtsova, I. G. (2013). On the absolute constants in the Berry–Esseen inequality and its structural and nonuniform improvements. *Informatika i Ee Primeneniya* **7** 124–125.

Sun, Q., Zhou, W.-X. and Fan, J. (2020). Adaptive Huber regression. *Journal of the American Statistical Association* **115** 254–265.

Tan, K. M., Wang, L. and Zhou, W.-X. (2022). High-dimensional quantile regression: Convolution smoothing and concave regularization. *Journal of the Royal Statistical Society: Series B* **84** 205–233.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B* **58** 267–288.

Volgushev, S., Chao, S.-K. and Cheng, G. (2019). Distributed inference for quantile regression models. *Annals of Statistics* **47** 1634–1662.

Wang, L. (2013). The L_1 regularized LAD estimator for high-dimensional linear regression. *Journal of Multivariate Analysis* **120** 135–151.

Wang, L. and He, X. (2021). Analysis of global and local optima of regularized quantile regression in high dimension: A subgradient approach. *Preprint*.

Wang, H., Li, G. and Jiang, G. (2007). Robust regression shrinkage and consistent variable selection through the LAD-Lasso. *Journal of Business and Economics Statistics* **25** 347–355.

Wang, J., Kolar, M., Srebro, N. and Zhang, T. (2017). Efficient distributed learning with sparsity. In *Proceedings of the 34th International Conference on Machine Learning* **70** 3636–3645.

Wang, H. J., Stefanski, L. A. and Zhu, Z. (2012). Corrected-loss estimation for quantile regression with covariate measurement errors. *Biometrika* **99** 405–421.

Wang, L., Wu, Y. and Li, R. (2012). Quantile regression for analyzing heterogeneity in ultra-high dimension. *Journal of the American Statistical Association* **107** 214–222.

Whang, Y.-J. (2006). Smoothed empirical likelihood methods for quantile regression models. *Econometric Theory* **22** 173–205.

Wu. T. T. and Lange, K. (2008). Coordinate descent algorithms for lasso penalized regression. *Annals of Applied Statistics* **2** 224–244.

Wu, Y., Ma, Y. and Yin, G. (2015). Smoothed and corrected score approach to censored quantile regression with measurement errors. *Journal of the American Statistical Association* **110** 1670–1683.

Yi, C. and Huang, J. (2017). Semismooth Newton coordinate descent algorithm for elastic-net penalized Huber loss regression and quantile regression. *Journal of Computational and Graphical Statistics* **26** 547–557.

Yu, Y., Chao, S.-K. and Cheng, G. (2020). Simultaneous inference for massive data: distributed bootstrap. In *Proceedings of the 37th International Conference on Machine Learning* **119** 10892–10901.

Yu, L., Lin, N. and Wang, L. (2017). A parallel algorithm for large-scale nonconvex penalized quantile regression. *Journal of Computational and Graphical Statistics* **26** 935–939.

Zhang, Y., Duchi, J. and Wainwright, M. (2015). Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates. *Journal of Machine Learning Research* **16**(102): 3299–3340.