# TOPICS IN MODERN REGRESSION MODELING

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Tao Zhang

August 2022

TOPICS IN MODERN REGRESSION MODELING

Tao Zhang, Ph.D.

Cornell University 2022

In the first part of this work, we propose a novel efficient sampling method for measurement-constrained data. Under "measurement constraints," responses are expensive to measure and initially unavailable on most of records in the dataset, but the covariates are available for the entire dataset. Our goal is to sample a relatively small portion of the dataset where the expensive responses will be measured and the resultant sampling estimator is statistically efficient. Measurement constraints require the sampling probabilities can only depend on a very small set of the responses. A sampling procedure that uses responses at most only on a small pilot sample will be called "response-free." We propose a response-free sampling procedure (OSUMC) for generalized linear models (GLMs). Using the A-optimality criterion, i.e., the trace of the asymptotic variance, the resultant estimator is statistically efficient within a class of sampling estimators. We establish the unconditional asymptotic distribution of a general class of response-free sampling estimators. This result is novel compared with the existing conditional results obtained by conditioning on both covariates and responses. Under our unconditional framework, the subsamples are no longer independent and new martingale techniques are developed for our asymptotic theory. We further derive the A-optimal response-free sampling distribution. Since this distribution depends on population level quantities, we propose the Optimal Sampling Under Measurement Constraints (OSUMC) algorithm to approximate the theoretical optimal sampling. Finally, we conduct an intensive

empirical study to demonstrate the advantages of OSUMC algorithm over existing methods in both statistical and computational perspectives. We find that OSUMC's performance is comparable to that of sampling algorithms that use complete responses. This shows that, provided an efficient algorithm such as OSUMC is used, there is little or no loss in accuracy due to the unavailability of responses because of measurement constraints.

In the second part of this work, we develop uniform inference methods for the conditional mode based on quantile regression. Specifically, we propose to estimate the conditional mode by minimizing the derivative of the estimated conditional quantile function defined by smoothing the linear quantile regression estimator, and develop two bootstrap methods, a novel pivotal bootstrap and the nonparametric bootstrap, for our conditional mode estimator. Building on high-dimensional Gaussian approximation techniques, we establish the validity of simultaneous confidence rectangles constructed from the two bootstrap methods for the conditional mode. We also extend the preceding analysis to the case where the dimension of the covariate vector is increasing with the sample size. Finally, we conduct simulation experiments and a real data analysis using U.S. wage data to demonstrate the finite sample performance of our inference method. The supplemental materials include the wage dataset, R codes and an appendix containing proofs of the main results, additional simulation results, discussion of model misspecification and quantile crossing, and additional details of the numerical implementation.

In the third part of this work, we develop a multi-round aggregated one-step estimator and a scalable bootstrap method for distributed sparse least absolute deviation (LAD) regression with high-dimensional covariates. The proposed one-step estimator is based on multi-round distributed quantile regression and

linear regression estimators. We derive convergence rates and sparsity properties of the new multi-round estimators and show that our multi-round one-step estimator requires less restrictive sample complexity than the one-shot aggregation for valid inference. We also develop a novel pivotal bootstrap for simultaneous inference that is scalable to the distributed setting. Building on high-dimensional Gaussian approximation techniques, we establish the validity of simultaneous confidence rectangles constructed from the pivotal bootstrap. Finally, we conduct numerical experiments using the simulated data, which demonstrate encouraging performance of our methods.

## BIOGRAPHICAL SKETCH

Tao Zhang was born in Hunan Province, China. After he finished high school at Changjun High School in Changsha, Hunan, he went to Fudan University in Shanghai where he obtained his bachelor degree in Mathematics and Applied Mathematics. He cultivated his interest in statistics during his undergraduate study and started his Ph.D. study in statistics at Cornell University right after his graduation from Fudan University. At Cornell, he conducted research on high-dimensional regression and inference, quantile regression and bootstrap inference and he was fortunate to be co-advised by Professor David Ruppert and Professor Kengo Kato.

To my parents and my girl friend, Peiying.

## ACKNOWLEDGEMENTS

First, I would like to express my sincere and deep gratitude to my advisors, Professor David Ruppert and Professor Kengo Kato, for their unselfish mentoring, guidance and support through out my Ph.D. career. Their devotion and rigorous attitude towards research and education has set them as my life-time role models of meticulous researchers. Thanks to their continuous support, endless patience and inspiring discussions, I was able to develop the spirit of curiosity, the skills of doing research and the rigorous attitude towards writing and presentation all of which are extremely important to my growth as an independent researcher and would be very beneficial in my future career. My deep gratitude also extends to Professor Yang Ning and Professor Marten Wegkamp for their insightful discussions and suggestions. I was fortunate to work with Professor Yang Ning on the first project of this dissertation and take classes taught by Professor Marten Wegkamp both of which are essential parts of my Ph.D. training.

Second, I would like to thank my family, especially my parents, for their unconditional and unselfish support and care which warms my heart during my down times and motivates me to keep moving forward. I also thank my girl friend, Peiying Song, for her valuable and consistent companion and encouragement.

Last but not least, I would like to thank my Ph.D. fellows and friends at Cornell for their support and the joy they brought to me. It is they who make my Ph.D. life at Cornell so colorful and rewarding.

## TABLE OF CONTENTS

## LIST OF TABLES

## LIST OF FIGURES

CHAPTER 1

# OPTIMAL SAMPLING FOR GENERALIZED LINEAR MODELS UNDER MEASUREMENT CONSTRAINTS

## 1.1 Introduction

### 1.1.1 Motivation and Contribution

Measurement constrained datasets (Wang et al., 2017c) where only a small portion of the data points have known a response, $Y$, but the covariates, $X$, are available for all data points, are common in practice. Datasets of this type can happen when the response is more expensive or time-consuming to collect than covariates. We will present two motivating examples. For more real-world examples, we refer readers to semi-supervised learning literature (e.g., Zhu 2005; Chapelle et al. 2010; Chakrabortty and Cai 2018).

**Example 1** (Critical Temperature of Superconductors)**.** Critical temperature, which is sensitive to chemical composition, is an important property of superconducting materials. Since no scientific model for critical temperature prediction is available (Hamidieh, 2018), a data-driven prediction model is desirable to guide researchers synthesizing superconducting materials with higher critical temperature. Due to the cost in both money and time for material synthesis, only a small portion of the thousands of potential chemical compounds can be manually tested. So selecting representative compositions to build a statistical model with maximum efficiency is important.

**Example 2** (Galaxy Classification)**.** Galaxy classification is an important task in astronomy (Banerji et al., 2010). Visual classification is time-consuming and

expensive and is becoming infeasible because the size of astronomical datasets is growing rapidly as more advanced telescopes enter operation. The size of modern galaxy datasets is often of millions or even billions (Reiman and Göhre, 2019). It is important to select a representative subsample of galaxies which can be classified accurately by humans, so that an effective classification model can be built.

In addition to many responses being missing, another characteristic of the datasets in these examples is the extreme size which brings huge challenges to statistical computing, data storage and communication. Sampling is a popular approach to super-large datasets where a small portion of the dataset is sampled and used as a surrogate of the entire dataset. There is a large literature on this problem, e.g., Wang et al. (2018), but many of the proposed sampling methods assume that the response is known on the entire dataset so are not applicable under measurement constraints. Moreover, they might not increase statistical efficiency even if they were applicable.

The problem addressed in the paper is fitting a generalized linear model (GLM) efficiently to massive measurement-constrained datasets using only a relatively-small subsample obtained by sampling. Measurement constrained datasets require sampling methods to be *response-free* meaning that the sampling probabilities can only depend on responses of a small pilot sample. Though a huge literature has been devoted to sampling methods for datasets where all records have both responses and covariates, efficient sampling algorithms under measurement constraints, which selects subsamples in a (nearly) response-independent way, are less well-studied. We bridge this gap by proposing a statistically efficient sampling method under measurement constraint (OSUMC)

for GLMs which does not require knowledge of responses (except in the pilot sample). Specifically, OSUMC selects subsamples with replacement according to a response-free A-optimal sampling distribution and computes the estimator by solving a weighted score equation based on the selected subsamples. We defer the details to Section 1.3 and Section 1.4.

This paper contributes to both theoretical and computational perspectives of the growing literature on statistical sampling. On the theoretical side, we prove asymptotic normality of a general class of response-free sampling estimators without conditioning on the data. Our asymptotic results are significantly different from conditional asymptotics in the traditional sampling literature which condition on both covariates and responses. In particular, the conditional independence of subsamples in the conditional asymptotics is no longer true under our framework. To deal with the correlation in our sampling estimator, we develop novel martingale techniques for our asymptotic theory. Another significant difference is that our asymptotic results compare our sampling estimator to the true parameter rather than to the maximum likelihood estimator as done by Wang et al. (2018). Since the true parameter is the object of interest, our results are more informative. Based on the asymptotic theory, we derive the optimal sampling probabilities by minimizing the A-criterion (Khuri et al., 2006) i.e., the trace of asymptotic variance matrix, which is equivalent to minimizing the sum of the asymptotic mean squared errors. On the computational side, we propose the OSUMC algorithm to approximate the theoretical A-optimal sampling method. Our sampling algorithm achieves an *optimal design* by assigning higher probabilities to data points that achieve estimators with higher efficiency. We show in our numerical study that the performance of OSUMC is comparable to that of sampling algorithms which use complete responses to calculate sam-

pling probabilities. This shows that, provided that an efficient algorithm such as OSUMC is used, there is little or no loss in accuracy due to not having the responses available for sampling because of measurement constraints.

### 1.1.2 Related Work

A large literature provide numerous variants of subsampling algorithms for linear regression (Drineas et al., 2006, 2011; Ma et al., 2015; Wang et al., 2019). One traditional approach is the leverage sampling which defines the sampling probabilities based on the empirical leverage scores of the design matrix (Huber, 2004). In more recent literature, leverage sampling serves as an important basis for more refined sampling methods of linear regression (Drineas et al., 2006, 2011; Ma et al., 2015). These papers fall in a paradigm termed *algorithmic leveraging*, which is fundamental to randomized numerical linear algebra (RandNLA) which aims at developing fast randomized algorithms for large-scale matrix-based problems. We refer readers to Ma et al. (2015) and Drineas and Mahoney (2016) for more references. However, most of the algorithmic leveraging literature is concerned with the numerical performance of algorithms and only a few papers, for example, Ma et al. (2015) and Raskutti and Mahoney (2016), provide statistical guarantees. Ma et al. (2015) derives bias and variance for their sampling methods while Raskutti and Mahoney (2016) provide statistical error bounds for their sketching method. In contrast, we will focus on the asymptotic efficiency which has not been considered in the algorithmic leveraging or RandNLA literature. Because leverage sampling for the least-square estimator is response-free it can adapt to measurement constraints, but this is not true for other GLMs.

For other GLMs, Wang et al. (2018) developed an A-optimal sampling procedure for logistic regression. Though a similar optimality criterion is adopted, our paper is very different from Wang et al. (2018). Most importantly, the sampling procedure in Wang et al. (2018) requires the complete set of responses and therefore cannot be used under measurement constraints, while OSUMC is tailored for that setting. From a theoretical perspective, the unconditional asymptotics derived in this paper are much more challenging than the conditional asymptotic setting considered in Wang et al. (2018), which conditions on both covariates and responses. The essential conditional independence assumption in the proof of Wang et al. (2018) is no longer valid and the correlated structure of the sampling estimator must be treated in our asymptotic theory. New martingale techniques are developed which are significantly different from the techniques used in Wang et al. (2018). Besides, unconditional framework allows us to prove the asymptotic results between our sampling estimator and the true parameter, instead of the conditional MLE as in Wang et al. (2018). A recent paper by Ai et al. (2018) generalized the results in Wang et al. (2018) to other GLMs under a similar framework, so again not applicable to the problem of measurement constraints. In another paper by Ting and Brochu (2018), the authors studied optimality of sampling for asymptotic linear estimators. Their conclusions will reduce to exactly the same results in Wang et al. (2018) for logistic regression, and hence cannot deal with measurement constraints neither.

One new research area also dealing with measurement constraints is *semi-supervised learning* (SSL). SSL attempts to use the unlabeled $X$ ($X$ without $Y$) to improve statistical performance. Though a huge SSL literature is devoted to algorithmic aspects (Zhu, 2005; Chapelle et al., 2010), only a few recent papers studied statistical estimation under the semi-supervised setting (Zhang

et al., 2016; Cai and Guo, 2018; Chakrabortty and Cai, 2018). For example, Chakrabortty and Cai (2018) estimate linear regression coefficients by regressing imputations of unobserved responses on the corresponding covariates. However, their method is computationally prohibitive for large datasets due to the nonparametric imputation approach while our method is computationally affordable even for massive datasets as demonstrated in out empirical study.

Another closely related area is *optimal experiment design* which determines the settings of covariates that yield estimators with optimal properties (Khuri et al., 2006). Whereas the design is determined freely in classical experiment design (Pukelsheim, 2006), the design points of this paper must be selected from the original dataset. Also, the solutions of the traditional optimal design problem are often combinatorial, which are computationally infeasible for even moderate size datasets. Wang et al. (2017c) propose sampling algorithms based on the convex relaxation of the traditional combinatorial problem. However, Wang et al. (2017c) mainly focus on algorithms and only a mean square error bound is provided for the statistical guarantee. Their results are also proved under traditional conditional framework which is different from the unconditional analysis of this paper. Additionally, independence assumptions for with-replacement sampling in Wang et al. (2017c) cannot be justified in many real situations, for instance, the two motivating examples. Such assumptions are avoided in our theory. On the practical side, our method offers simple closed forms of the optimal sampling weights while Wang et al. (2017c)'s procedure involves solving a semi-definite programming problem which can be computationally intensive for large datasets.

## 1.2 Background and Setup

We begin with the background on GLMs. Assume $n$ independent and identically distributed data couples, $(X_1, Y_1), \ldots, (X_n, Y_n) \sim (X, Y)$, where $X \in \mathbb{R}^p$ is a covariate vector, $Y \in \mathbb{R}$ is the response, and $Y$ given $X$ satisfies a GLM with the canonical link:

$$P(Y|X, \beta_0, \sigma) \propto \exp\left\{\frac{Y \cdot X^T\beta_0 - b(X^T\beta_0)}{c(\sigma)}\right\},$$

Here $b(\cdot)$ is a known function, $\sigma$ is a known dispersion parameter, and $\beta_0 \in \mathbb{R}^p$ is the unknown parameter of interest and assumed to be in a compact set $\mathcal{B} \subseteq \mathbb{R}^p$. Without loss of generality, we take $c(\sigma) = 1$. The standard estimator of $\beta_0$ is the MLE

$$\hat{\beta}_{\text{mle}} \in \arg\min_{\beta \in \mathbb{R}^p} \left[-\frac{1}{n}\sum_{i=1}^{n}\left\{Y_i \cdot X_i^T\beta - b(X_i^T\beta)\right\}\right].$$

Equivalently, we could solve the score equation to obtain the MLE

$$\Psi_n(\beta) := \frac{1}{n}\sum_{i=1}^{n}\psi_\beta(X_i, Y_i) = \frac{1}{n}\sum_{i=1}^{n}\left\{b'(X_i^T\beta) - Y_i\right\} \cdot X_i = 0.$$

Iterative methods such as Newton's method and its variants are usually adopted to solve such problems numerically (McCullagh and Nelder, 1989; Aitkin et al., 2005). If the sample size $n$ is very large, the computational cost for just one iteration will be huge. Therefore, a sampling approach can be used to reduce computational cost.

In addition, we assume a measurement constraint setting where only a small portion of responses are available initially. As mentioned in the introduction, this setting is common in practice. The main purpose of this paper is to develop a unified and statistically efficient sampling procedure for GLMs under measurement constraints.

## 1.3 General Sampling Scheme

We first present in Algorithm 1 a general response-free sampling scheme for GLMs. The class of response-free sampling estimators is defined accordingly.

**Algorithm 1:** Response-free sampling procedure for GLMs

1. Sample with replacement from the original $n$ data points $r$ times with probabilities $\pi = \{\pi_i\}_{i=1}^n$, where we require that $\pi_i$ only depends on $(X_1, \ldots, X_n)$ and a pilot estimate of $\beta$, but not $(Y_1, \ldots, Y_n)$. Collect the subsample $(X_i^*, Y_i^*)_{i=1}^r$, where we let $(X_i^*, Y_i^*)$ denote the data sampled out in the $i$-th step.
2. Define the reweighted score function as

$$\Psi_n^*(\beta) := \frac{1}{r} \sum_{i=1}^{r} \frac{b'(X_i^{*T}\beta) - Y_i^*}{n\pi_i^*} \cdot X_i^*$$

   where $\pi_i^*$ corresponds to the sampling probability of $(X_i^*, Y_i^*)$.
3. Solve the reweighted score equation $\Psi_n^*(\beta) = 0$ to get the estimator $\hat{\beta}_n$.

We emphasize that Algorithm 1 is tailored for the measurement constraints setting, as the sampling weight $\pi_i$ only depends on $(X_1, \ldots, X_n)$ not $(Y_1, \ldots, Y_n)$ which may not be completely observable. In practice, once the subsample of size $r$ are selected, one need measure only those $r$ responses. This may bring huge economic savings as response measurements are usually expensive in measurement constrained situations.

An additional benefit of sampling is cost savings. In the last step of Algorithm 1, Newton's method and its equivalent variants are usually adopted (McCullagh and Nelder, 1989; Aitkin et al., 2005). Given the sampling probabilities $\pi$, Algorithm 1 dramatically reduces the computational and storage costs by making them scale in $r$ instead of $n$ which could be much larger than $r$. More

concretely, if $n = 10^6$ and $p = 20$, the computational time of one iteration of Newton's method for the full sample MLE is $O(np^2) = O(4 \times 10^8)$. In addition, if each data point occupies 1MB of storage space, then the original dataset would occupy around 1TB space. In contrast, for Algorithm 1 with $r = 1,000$, the computational time for each iteration is $O(4 \times 10^5)$ and the storage space is less than 1 GB, which substantially lower the computational and storage cost.

The performance of Algorithm 1 depends crucially on the choice of the sampling weights $\pi_i$ and the subsample size $r$. Under measurement constraints, subsample size $r$ is determined by the cost of measuring the response and, perhaps, by the availability of the computational or storage resources. The more important question is how to determine the sampling distribution in a data-driven approach such that the resultant sampling estimator achieves optimal efficiency. We will answer this question in the next section by defining the A-optimal sampling distribution based on the asymptotic results therein.

## 1.4 Optimal Sampling Procedure and Asymptotic Theory

In this section, we assume the classical asymptotic setting in which $n \to \infty$ and $p$ is fixed. We first show the asymptotic normality of the sampling estimator defined in Algorithm 1, and then find the A-optimal sampling distribution by minimizing the asymptotic mean squared error.

### 1.4.1 Notation

The $j$th entry of the covariate vector $X_i$ is denoted by $x_{ij}$. For $X \in \mathbb{R}^p$, $||X||$ is the Euclidean norm of $X$. We also define tuple notations: $X_1^n := (X_1, X_2, \ldots, X_n)$ and $Y_1^n := (Y_1, Y_2, \ldots, Y_n)$. $V(X)$ and $E(X)$ denote the variance and expectation of $X$, respectively.

### 1.4.2 Consistency of $\hat{\beta}_n$

We first show the statistical consistency of $\hat{\beta}_n$.

**Theorem 1** (Consistency of $\hat{\beta}_n$)**.** *Assume the following conditions*

*(i)* *Either* $(ia)$ $b''(\cdot)$ *is bounded or* $(ib)$ $X$ *is bounded.*

*(ii)* $EXX^T$ *is finite and* $\zeta(\beta) := E\left[\{b'(X^T\beta) - Y\}X\right]$ *is finite for any* $\beta \in \mathcal{B}$.

*(iii)* $\sum_{i=1}^n E\left[\frac{\{b'(X_i^T\beta)-Y_i\}^2}{\pi_i} x_{ij}^2\right] = o(n^2 r)$ *for* $1 \leq j \leq p$ *and* $\beta \in \mathcal{B}$.

*(iv)* $\inf_{\beta:||\beta-\beta_0||\geq\epsilon} ||\zeta(\beta)|| > 0$ *for any* $\epsilon > 0$.

*Then* $\hat{\beta}_n \xrightarrow{p} \beta_0$.

Condition (i$a$) is satisfied for most of GLMs except for Poisson regression. For Poisson regression, our theory is still applicable if Condition (ib) is satisfied. Condition (iii) - (iv) are needed to apply a consistency theorem for M-estimators (van der Vaart, 2000). In particular, condition (iii) ensures the uniform convergence of $\Psi_n^*$ while condition (iv) is a common *well-separated* condition for consistency proofs which is satisfied if $\zeta(\cdot)$ has unique minimizer and if, as we have assumed, the parameter space$\mathcal{B}$ is compact.

### 1.4.3 Asymptotic Normality of $\hat{\beta}_n$

To establish asymptotic normality of $\hat{\beta}_n$, we start with an important asymptotic representation.

**Lemma 1** (Asymptotic linearity). *Assume the following conditions,*

*(i)* $\Phi = E\left\{b''(X^T\beta_0)XX^T\right\}$ *is finite and non-singular.*

*(ii)* $\sum_{i=1}^{n} E\left\{\frac{b''(X_i^T\beta_0)^2}{\pi_i}(x_{ik}x_{ij})^2\right\} = o(n^2 r)$, *for* $1 \leq k, j \leq p$.

*(iii)* $b(x)$ *is three-times continuously differentiable for every* $x$ *within its domain.*

*(iv) Every second-order partial derivative of* $\psi_\beta(x)$ *w.r.t* $\beta$ *is dominated by an integrable function* $\ddot{\psi}(x)$ *independent of* $\beta$ *in a neighborhood of* $\beta_0$.

*If* $\Psi_n^*(\hat{\beta}_n) = 0$ *for all large* $n$ *and if* $\hat{\beta}_n$ *is consistent for* $\beta_0$, *then*

$$\Psi_n^*(\beta_0) = -\Phi(\hat{\beta}_n - \beta_0) + o_p(||\hat{\beta}_n - \beta_0||).$$

The proof of Lemma 1 is based on the asymptotic linearity of M-estimators, e.g., Theorem 5.41 in van der Vaart (2000). It turns out to be important for the establishment of asymptotic normality.

We will apply a multivariate martingale central limit theorem (Lemma 4 in the supplementary materials) to the above asymptotic linear representation and show the asymptotic normality of $\hat{\beta}_n$.

We first define a filtration $\{\mathcal{F}_{n,i}\}_{i=1}^{r(n)}$ adaptive to our sampling procedure: $\mathcal{F}_{n,0} = \sigma(X_1^n, Y_1^n)$; $\mathcal{F}_{n,1} = \sigma(X_1^n, Y_1^n) \vee \sigma(*_1)$; $\cdots$; $\mathcal{F}_{n,i} = \sigma(X_1^n, Y_1^n) \vee \sigma(*_1) \vee \cdots \vee \sigma(*_i)$; $\cdots$, where $\sigma(*_i)$ is the $\sigma$-algebra generated by $i$th sampling step, which can be interpreted as the smallest $\sigma$-algebra that contains all the information in

$i$th step. In the following, we always assume subsample size $r$ is increasing with $n$. Based on the filtration, we define the martingale

$$M := \sum_{i=1}^{r} M_i := \sum_{i=1}^{r} \left[ \frac{b'(X_i^{*T}\beta_0) - Y_i^*}{rn\pi_i^*} \cdot X_i^* - \frac{1}{rn} \sum_{j=1}^{n} \{b'(X_j^T\beta_0) - Y_j\} \cdot X_j \right],$$

where $\{M_i\}_{i=1}^r$ is a martingale difference sequence adapt to $\{\mathcal{F}_{n,i}\}_{i=1}^r$. In addition, we define: $Q := \frac{1}{n}\sum_{j=1}^{n}(b'(X_j^T\beta_0) - Y_j) \cdot X_j$; $T := \Psi_n^*(\beta_0) = M + Q$; $\xi_{ni} := V(T)^{-\frac{1}{2}} M_i$; $B_n := V(T)^{-\frac{1}{2}} V(M) V(T)^{-\frac{1}{2}}$, which is the variance of the normalized martingale $V(T)^{-\frac{1}{2}} M$.

**Theorem 2** (Asymptotic normality of $\hat{\beta}_n$)**.** *Under the conditions in Lemma 1 and we further assume*

*(i)* $\lim\limits_{n\to\infty} \sum\limits_{i=1}^{r} E[||\xi_{ni}||^4] = 0$,

*(ii)* $\lim\limits_{n\to\infty} E[|| \sum\limits_{i=1}^{r} E[\xi_{ni}\xi_{ni}^T | \mathcal{F}_{n,i-1}] - B_n||^2] = 0$,

*we have*

$$V(T)^{-\frac{1}{2}} \Phi(\hat{\beta}_n - \beta_0) \xrightarrow{d} N(0, I).$$

Condition (i) and (ii) are martingale version integrability conditions similar to Lindeberg-Feller conditions. In fact, such conditions are common in martingale central limit theorems (Hall and Heyde, 2014).

### 1.4.4 Optimal Sampling Weights under Measurement Constraints

In this section, we will derive the A-optimal sampling distribution for our general response-free sampling procedure.

Theorem 2 shows that for sufficiently large subsample size, the distribution of $\hat{\beta}_n - \beta_0$ can be well approximated by $N(0, \mathbf{V})$ with $\mathbf{V} := \Phi^{-1}V(T)\Phi^{-1}$. If $\beta_0$ is univariate, we can optimize the sampling probability by minimizing the asymptotic variance $\Phi^{-1}V(T)\Phi^{-1}$ which results in highest statistical efficiency. For multi-dimensional $\beta_0$, we adopt the A-optimality criterion of experiment design (Kiefer, 1959) and minimize the trace of the covariance matrix. Minimization of the trace of $\mathbf{V}$, i.e. $\mathrm{tr}(\Phi^{-1}V(T)\Phi^{-1})$, is equivalent to minimization of the asymptotic mean squared error. The following theorem specifies A-optimal sampling distribution.

**Theorem 3.** *If for $1 \leq j \leq n$, the sampling probability is*

$$\pi_j = \frac{\sqrt{b''(X_j^T\beta_0)}||\Phi^{-1}X_j||}{\sum_{i=1}^n \sqrt{b''(X_j^T\beta_0)}||\Phi^{-1}X_j||},$$

*then $\mathrm{tr}(\mathbf{V})$ will attain its minimum, i.e., $\{\pi_j\}_{j=1}^n$ is the A-optimal sampling distribution.*

The optimal weights cannot be calculated directly in practice, since they depend on population level quantities $\Phi^{-1}$ and $\beta_0$. Therefore, to implement response-free sampling, we need pilot estimates of $\Phi$ and $\beta_0$. The details are shown in Algorithm 2.

**Remarks**

- Step 1 in Algorithm 2 is designed for pilot estimation when very few or even none of the responses are available initially, but expensive responses collection is possible. If a moderate number of responses are accessible in the initial data pool, we could use them to calculate pilot estimators. Otherwise, a small random sample can be taken with uniform sampling

**Algorithm 2:** Optimal Sampling under Measurement Constraint (OSUMC)

1. Uniformly sample $r_0$ ($\ll r$) data points with indices $i_1, \ldots, i_{r_0}$ and collect those points: $\{(X_{i_j}, Y_{i_j})\}_{j=1}^{r_0}$ from data pool. Calculate $\tilde{\beta}_n$, the pilot estimator of $\beta_0$, and $\tilde{\Phi}_n := \frac{1}{r_0} \sum_{j=1}^{r_0} b''(X_{i_j}^T \tilde{\beta}_n) X_{i_j} X_{i_j}^T$, the pilot estimator of $\Phi$, based on the $r_0$ data points.
2. Calculate the approximate optimal sampling weight for each data point:

$$\pi_j \propto \sqrt{b''(X_j^T \tilde{\beta}_n)} ||\tilde{\Phi}_n^{-1} X_j||,$$

   for $1 \leq j \leq n$.
3. Run Algorithm 1 with $\pi_j$ defined above to obtain the final estimator $\hat{\beta}_n$

probabilities. The size of the pilot sample, $r_0$ could be fairly small compared with total sample size $n$ and even of the same order of magnitude as the dimension $p$ of the problem. In our empirical study, we set $r_0 = 500$ with the complete dataset of size $n = 10^5$ and $p$ ranging from $20$ to $100$, and the performance of Algorithm 2 is satisfactory.

- (Computational complexity) When we use Newton's method, or one of its variants such as Fisher scoring, to compute the root of the score equation, the computation requires $O(\zeta np^2)$ computational time, where $\zeta$ is the number of iterations needed for the algorithm to converge. In our empirical study, $\zeta$ varies from $10$ to $30$ under different models. For OSUMC algorithm the first step requires $O(\zeta_1 r_0 p^2)$ computation time where $\zeta_1$ is the number of iterations. In the second step, $O(np^2 + \zeta_2 rp^2)$ computation time is required where $\zeta_2$ is the number of the iterations in this step. Hence, OSUMC algorithm has complexity of order $O(np^2 + \zeta_1 r_0 p^2 + \zeta_2 rp^2)$. If $n$ is extremely large such that $p$, $r_0$, $r$, $\zeta_1$ and $\zeta_2$ are all much smaller than $n$, the computation complexity of the algorithm is $O(np^2)$. Therefore,

the computational advantage of using OSUMC algorithm compared with full-sample MLE is still huge if the scale of the problem is large, i.e., $np^2$ is large and $\zeta > 10$. The intensive numerical study in the following provides strong evidence for such advantage. To further reduce the computational complexity, one may use modified Newton's method for large-scale computation; see, for example, Xu et al. (2016).

## 1.5 Numerical Examples

### 1.5.1 Simulation Results

In this section, we evaluate the performance of the OSUMC algorithm on synthetic datasets. Due to page limitation, we will show the numerical results for logistic and linear regression and defer Poisson regression to Section A.6 in the supplementary material. All the results are obtained in the R environment with one Intel Xeon processor and 8 Gbytes RAM over Red Hat OpenStack Platform.

#### Logistic Regression

We generate datasets of size $n = 100,000$ from the logistic regression model,

$$P(Y = 1|X, \beta_0) = \frac{\exp(X^T\beta_0)}{1 + \exp(X^T\beta_0)},$$

where $\beta_0$ is a $20$ dimensional vector with all entries $1$. We consider four different scenarios to generate $X$ as in Wang et al. (2018).

- **mzNormal.** $X$ follows the multivariate normal distribution $N(0, \Sigma)$ with

$\Sigma_{ij} = 0.5^{I(i \neq j)}$. In this case, we have a balanced dataset, i.e., the number of 1's and the number of 0's in the responses are almost equal.

- **nzNormal.** $X$ follows the multivariate normal distribution $N(0.5, \Sigma)$. In this case, we have an imbalanced dataset where about $75\%$ of the responses are 1's.
- **unNormal.** $X$ follows the multivariate normal distribution with mean zero but different variances. To be more specific, $X$ follows the multivariate normal distribution $N(0, \Sigma_1)$ with $\Sigma_1 = U_1 \Sigma U_1$, where $U_1 = diag(1, 1/2, \ldots, 1/20)$.
- **mixNormal** $X \sim 0.5N(0.5, \Sigma) + 0.5N(-0.5, \Sigma)$.

In each case, we compare our optimal sampling procedure (OSUMC) with the mMSE method in Wang et al. (2018) (OSMAC), uniform sampling (Unif), and the benchmark full data MLE under different subsample sizes ($r$). In our procedure, we set the subsample size of uniform sampling in the first step equal to $r_0 = 500$. For uniform sampling, we directly subsample $r$ points and calculate the subsample MLE.

We repeat simulations $S = 500$ times, and calculate the empirical MSE as $S^{-1} \sum_{s=1}^{S} ||\hat{\beta}_n^{(s)} - \beta_0||^2$ where $\hat{\beta}_n^{(s)}$ is the estimate from the $s$th repetition. The comparison of the empirical MSE is presented in Figure 1.1.

From Figure 1.1, both OSUMC method and the OSMAC method in Wang et al. (2018) uniformly dominate the uniform sampling method in all four scenarios, which agrees with Theorem 3. In most of the simulation settings (except for unNormal), our sampling procedure performs similarly to the OSMAC in Wang et al. (2018). This is because both methods adopt the A-optimality cri-

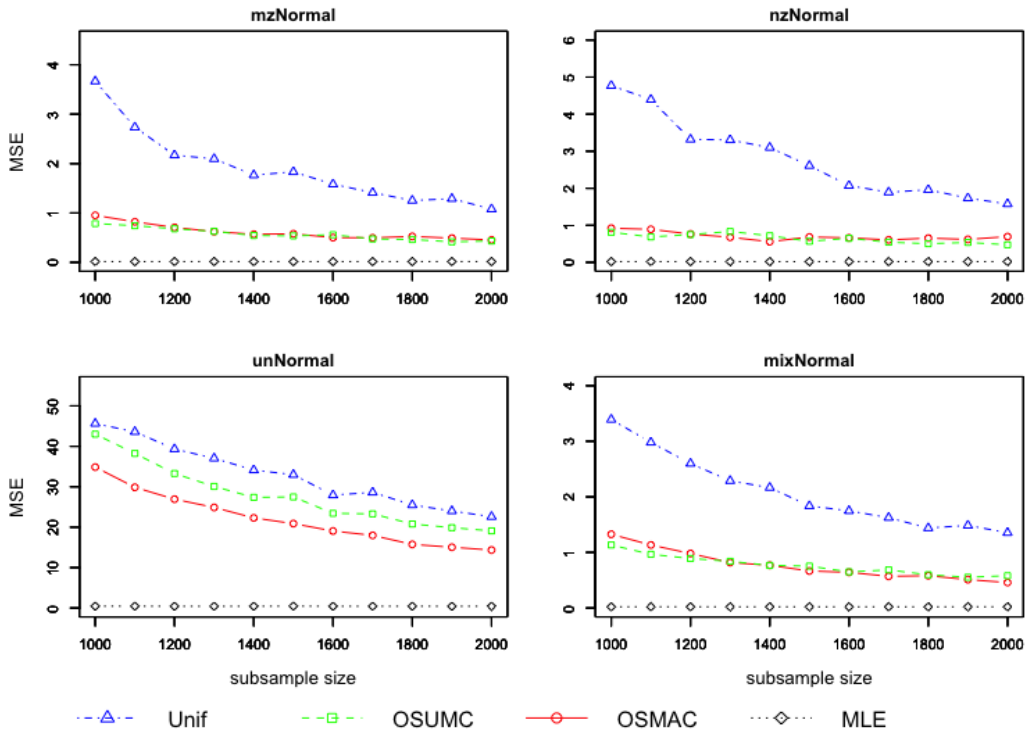



Figure 1.1: MSE of the proposed optimal sampling procedure (OSUMC), the method in Wang et al. (2018) (OSMAC), the uniform sampling (Unif), and the full sample MLE (MLE) for different subsample size $r$ under four scenarios in logistic regression.

terion in respective framework to derive the sampling weights. However, we note that OSMAC requires the response of each data point, which is infeasible under measurement constraints, while our method can be implemented as long as a moderate number of responses are available for the pilot estimators.

We also compare the average computational time for each method under all scenarios and the computational time plot can be found in Section A.5.1 in the supplementary materials. Our simulation reveals that the computation time is not very sensitive to the subsample size for all the four methods. In most cases, OSUMC and OSMAC perform similarly and require significantly less computational time compared with the full-sample MLE.

To see whether the asymptotic normality in our theory holds under the previous four different design generation settings, we plot the chi-square Q-Q plot of the resultant estimator $\hat{\beta}_n$ based on $1000$ simulations with fixed subsample size $r = 5000$ for each considered setting. The plots are presented in the supplementary materials, Section A.5.2. Q-Q plots reveal that the resultant sampling estimator $\hat{\beta}_n$ is approximately normal with sufficiently large sample size $n$ and subsample size $r$ in the four considered design generation settings.

**Linear Regression**

We generate datasets of size $n = 100,000$ and dimension $p = 100$ from the following linear regression model: $Y = X\beta_0 + \epsilon$, where $\beta_0 = (\underbrace{0.1, \ldots, 0.1}_{5}, \underbrace{10, \ldots, 10}_{90}, \underbrace{0.1, \ldots, 0.1}_{5})^T$ and $\epsilon \sim N(0, 9I_n)$.

We note that in linear regression model, OSUMC algorithm is equivalent to the following Algorithm 3. Algorithm 3 is similar to the general least-squares sampling meta-algorithm in Ma et al. (2015), which is adopted in Drineas et al. (2006, 2011, 2012). We consider the following design generation settings from Ma et al. (2015). Similar settings are also investigated in Wang et al. (2017c).

1. **GA.** The $n \times p$ design matrix $\mathbf{X}$ is generated from multivariate normal $N(1_p, \Sigma_2)$ with $\Sigma_2 = U_2 \Sigma U_2$, where $U_2 = \mathrm{diag}(5, 5/2, \ldots, 5/30)$.
2. $\mathbf{T_3}$. Design matrix $\mathbf{X}$ is generated from multivariate t-distribution with 3 degrees of freedom and covariance $\Sigma_2$ as GA.
3. $\mathbf{T_1}$. Design matrix $\mathbf{X}$ is generated from multivariate t-distribution with 1 degrees of freedom and covariance $\Sigma_2$ as GA.

**Algorithm 3:** Optimal Sampling for Linear Regression

1. Uniformly sample $r_0(\ll r)$ data points: $\{(X_{i_j}, Y_{i_j})\}_{j=1}^{r_0}$ . Calculate $\tilde{\Phi}_n := \frac{1}{r_0}\sum_{j=1}^{r_0} X_{i_j}X_{i_j}^T$, the pilot estimator of $\Phi$.
2. Calculate the approximate optimal sampling weight for each data point:

$$\pi_j \propto ||\tilde{\Phi}_n^{-1}X_j||,$$

   for $1 \leq j \leq n$.
3. Repeat sampling $r$ times according to probability in step $2$ and rescale each sampled data point $(X_i^*, Y_i^*)$ by $1/\sqrt{\pi_i^*}$, $1 \leq i \leq r$.
4. Calculate the ordinary least-squares estimator of the rescaled subsample and output it as the final estimator.

We compare our method (OSUMC) with uniform sampling (Unif), leverage sampling (Leverage) and shrinkage leveraging (SLEV) in Ma et al. (2015) over different subsample size in the three design generation settings above. For the SLEV method, the shrinkage parameter $\alpha$ is set to be $0.9$ as in Ma et al. (2015). Again, we repeat the simulation $500$ times and report the empirical MSE and computational time in Figures 1.2 and A.3, respectively.

For all three design generation settings, our method always results in smaller MSE than the other three methods, which again is consistent with our theoretical results. The advantage of our method becomes more obvious when the design generation distribution is more heavy-tailed by noting that $GA$ has moments of arbitrary orders while $T_k$ only has moments up to order $k-1$. It is interesting to see that our method outperforms other methods even in the $T_1$ design setting where the moment assumptions imposed in our theory are violated. The performance of both leverage sampling and shrinkage leverage sampling improves with heavier-tailed design generation distributions. This has

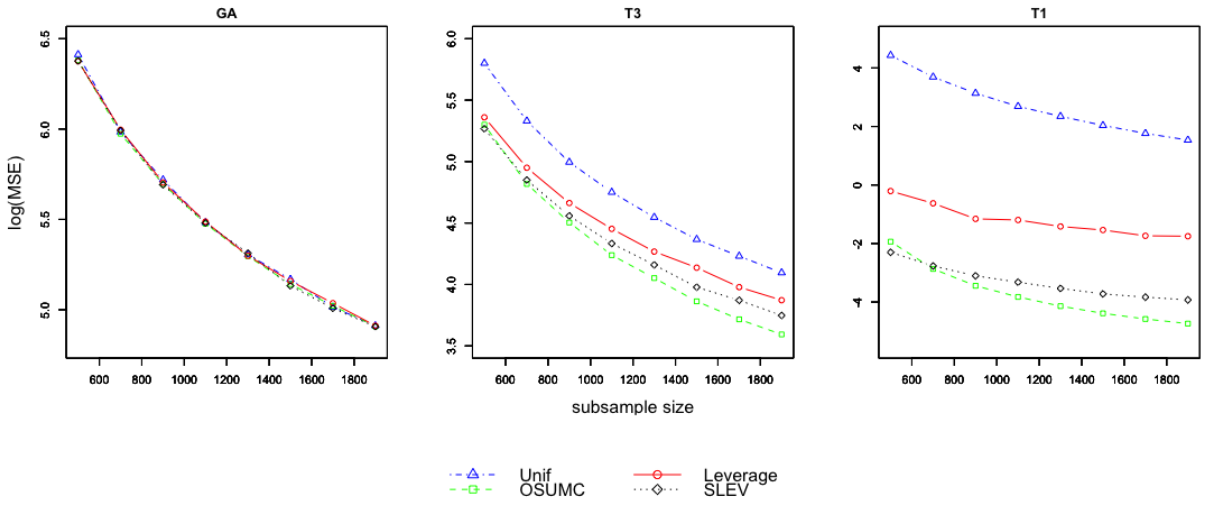



Figure 1.2: MSE plots for different subsample size $r$ under different design generation settings for linear regression

been well understood in the literature on leverage sampling and outlier diagnosis (Rousseeuw and Hubert, 2011; Ma et al., 2015). As expected, all the three sampling methods above yield smaller MSEs than uniform sampling.

The average computational times of the four methods are reported in Figure A.3 in the supplementary materials. Again, the results show the insensitivity of the computational time to increasing subsample sizes. Our method requires the second smallest computing time, being inferior only to the uniform sampling. Both leverage related methods take more than double the computational time of our method due to the intensive computation of leveraging score of each data point.

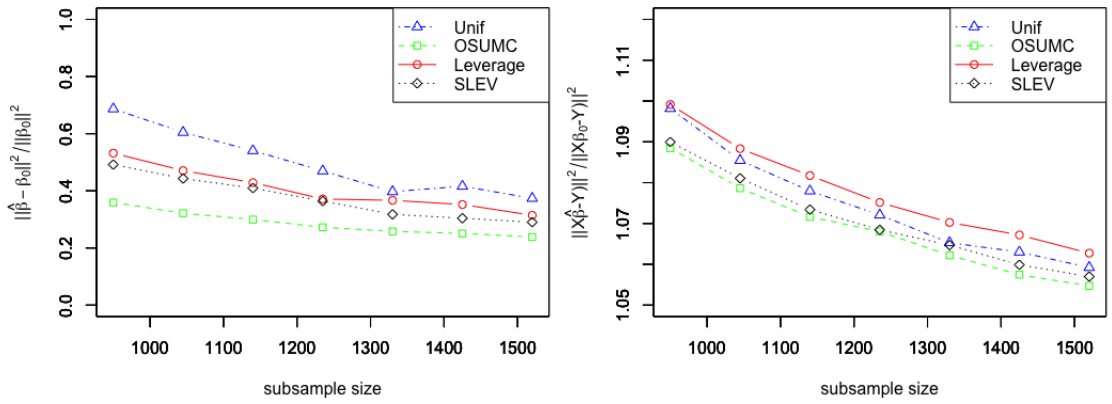
Again, we provide Q-Q plots of the resultant estimator $\hat{\beta}_n$ for each considered design setting and the results can be found in Section A.5.4 in the supplementary materials.

### 1.5.2 Superconductivty Dataset

In this section, we analyze the superconductivty dataset (Hamidieh, 2018), which is available from the Machine Learning Repository at: `https://archive.ics.uci.edu/ml/datasets/Superconductivty+Data#`. The purpose of Hamidieh (2018) is to build a statistical model to predict the superconducting critical temperature of superconducting materials based on their chemical formulas. In the dataset, 21,263 different superconductors' critical temperatures are collected along with $81$ features extracted from the chemical formulas the superconductors. Multiple linear model is considered in Hamidieh (2018) and regression coefficients are calculated based on the full sample, which is treated as "true" parameters ($\beta_0$) in the following analysis.

We compare our sampling method (OSUMC) with the three other sampling methods in the simulation studies for linear models. Besides the estimation accuracy which is the main focus before, prediction performance of the sampling algorithm will also be evaluated. We randomly select 19,000 data points as the training set and use the rest as the test set for prediction purpose. Then we implement the sampling method on the training dataset and obtain the coefficient estimator $\hat{\beta}$. We now measure the estimation and prediction performance by *relative mean squared error*: $||\hat{\beta} - \beta_0||^2/||\beta_0||^2$ and *prediction relative squared error*: $||X\hat{\beta} - Y||^2/||X\beta_0 - Y||^2$ which is calculated over test dataset, respectively. We repeat the process $500$ times for different subsample sizes and the median of the two criteria are recorded for each subsample size. The results are presented in Figure 1.3. We also report the median running times of the four sampling methods over different subsample sizes in Figure A.5.

Figure 1.3 shows that our method consistently achieves the best performance




(a) Estimation relative SE (b) Prediction relative SE

Figure 1.3: Estimation and prediction performance comparison of four different sampling methods over different subsample size

on both estimation and prediction. In addition, our sampling procedure outperforms both leverage-based methods in computational time. Though SLEV method achieves similar prediction accuracy as our method, it takes more than twice the computational time. Therefore, the proposed sampling procedure maintains a good balance between statistical efficiency and computational cost.

## 1.6 Summary

In this paper, we propose a novel sampling procedure, OSUMC, to address measurement constraints when estimating GLMs. We show unconditional asymptotic normality of a general class of response-free sampling estimators by using newly developed martingale techniques. Our unconditional asymptotic results obtained without conditioning on the data are different from existing conditional results which condition on both covariates and responses. Building on the asymptotic theory, we derive the A-optimal sampling distribution, which

depends on population level quantities. For practical applications, we propose OSUMC algorithm to approximate the theoretical optimal sampling scheme. Additionally, we conduct extensive numerical studies which show that the performance of OSUMC is comparable to that of sampling algorithms which use complete responses to calculate sampling probabilities. This indicates that OSUMC successfully prevents the loss of statistical efficiency due to measurement constraints.

A number of interesting extensions and open areas remain. For example, it would be interesting to consider the high-dimensional scenario where the dimension $p$ could be much larger than the subsample size $r$. Under such setting, techniques like regularization and debiasing in the high-dimensional statistics would be need which is out of the scope of this paper. We will leave it for future investigations.

CHAPTER 2

# BOOTSTRAP INFERENCE FOR QUANTILE-BASED MODAL REGRESSION

## 2.1 Introduction

### 2.1.1 Overview

Modal regression is a principal statistical methodology to estimate and make inference on the conditional mode. Modes provide useful distributional information missed by the mean when the (conditional) distribution is skewed (Chen et al., 2016) and are known to be robust under measurement errors (Bound and Krueger 1991; Hu and Schennach 2008). The global mode offers intuitive interpretability by being understood as "the most likely" or "the most common" (Heckman et al. 2001; Hedges and Shah 2003). As such, modal regression has wide applications in various areas including astronomy (Bamford et al., 2008), medical research (Wang et al., 2017b), econometrics (Kemp and Santos-Silva, 2012), etc. We refer the reader to Chacón (2018) and Chen (2018) for recent reviews on modal regression; see also a literature review below.

In this paper, we consider estimating the conditional mode by "inverting" a quantile regression model, which builds on the observation that the derivative of the conditional quantile function coincides with the reciprocal of the conditional density so that the conditional mode can be obtained by minimizing the derivative of the conditional quantile function. Specifically, we estimate the conditional mode by minimizing the derivative of the kernel smoothed

Koenker-Bassett estimator of the conditional quantile function (Koenker and Bassett, 1978) with a sufficiently smooth kernel. We develop asymptotic theory for the proposed estimator $\hat{m}(\boldsymbol{x})$ of the conditional mode $m(\boldsymbol{x})$. In particular, we consider simultaneous confidence intervals for the conditional mode at multiple design points, $m(\boldsymbol{x}_1), \ldots, m(\boldsymbol{x}_L)$, where $L$ is allowed to grow with the sample size $n$, i.e., $L = L_n \to \infty$. To this end, we first show that $\hat{m}(\boldsymbol{x}) - m(\boldsymbol{x})$ can be approximated by the linear term $(nh^{3/2})^{-1} \sum_{i=1}^n \psi_{\boldsymbol{x}}(U_i, \boldsymbol{X}_i)$ uniformly over a range of design points $\boldsymbol{x}$, where $h = h_n \to 0$ is a sequence of bandwidths, $\psi_{\boldsymbol{x}}$ is the influence function (that depends on $n$) at design point $\boldsymbol{x}$, $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are independent covariate vectors, and $U_1, \ldots, U_n$ are mutually independent uniform random variables on $(0, 1)$ independent of the covariate vectors. Building on high dimensional Gaussian approximation techniques developed in Chernozhukov et al. (2014, 2017a), we show that $\sqrt{nh^3}(\hat{m}(\boldsymbol{x}_\ell) - m(\boldsymbol{x}_\ell))_{\ell=1}^L$ can be approximated by an $L$-dimensional Gaussian vector uniformly over the hyperrectangles in $\mathbb{R}^L$, i.e., all sets $A$ of the form: $A = \{w \in \mathbb{R}^L \ : \ a_j \leq w_j \leq b_j \text{ for all } j = 1, \ldots, L\}$ for some $-\infty \leq a_j \leq b_j \leq \infty$, $j = 1, \ldots, L$, even when $L \gg n$.

As the limiting Gaussian distribution is infeasible in practice, we consider two bootstrap methods, the nonparametric bootstrap and a novel pivotal bootstrap, to conduct valid inference. We first discuss the motivation of the new pivotal bootstrap. The leading stochastic term in the prescribed expansion is conditionally "pivotal" in the sense that conditionally on $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, the distribution of the process

$$\boldsymbol{x} \mapsto (nh^{3/2})^{-1} \sum_{i=1}^n \psi_{\boldsymbol{x}}(U_i, \boldsymbol{X}_i)$$

is completely known up to some nuisance parameters. This suggests a version of bootstrap for the proposed estimator by sampling uniform random variables $U_i$ independent of the data. In practice, the influence function $\psi_{\boldsymbol{x}}$ depends

on nuisance parameters and we replace them by consistent estimates. We call the resulting bootstrap "pivotal bootstrap" and prove that the pivotal bootstrap can consistently estimate the sampling distribution of $\sqrt{nh^3}(\hat{m}(\boldsymbol{x}_\ell) - m(\boldsymbol{x}_\ell))_{\ell=1}^L$ uniformly over the rectangles in $\mathbb{R}^L$ even when $L \gg n$. In fact, our inference framework is more general and covers simultaneous inference for linear combinations of the vector $(m(\boldsymbol{x}_\ell))_{\ell=1}^L$, which can be used to construct simultaneous confidence intervals for partial effects and test significance of certain covariates on the conditional mode. We also establish a similar consistency result for the nonparametric bootstrap. Finally, we extend the previous analysis to the case where the dimension of the covariate vector increases with the sample size.

We conduct simulation experiments on various mode inference problems and a real data analysis to demonstrate the finite sample performance of the bootstrap methods. Our simulation experiments show that the pivotal bootstrap yields accurate pointwise and simultaneous confidence intervals for the conditional modes. Additionally, we apply our inference method to analyze a real U.S. wage dataset. Analysis of wage data is important in econometric and social science (Autor et al. 2008; Western and Rosenfeld 2011; Buchinsky 1994). Wage data are often positively skewed and "the most common wage" as a representative of the majority of the population is usually of more interest. Common questions in the analysis of wage data include: i) What is the most likely wage for given covariates? How to construct pointwise and simultaneous confidence intervals for the estimated wages? ii) Is there an effect of a specific covariate on the most likely wage given the same other covariates? We address those empirical questions using the inference method developed in the present paper.

From a technical perspective, the asymptotic analysis in this paper is highly

nontrivial. Our program of the technical analysis proceeds as 1) first establishing a uniform asymptotic representation and 2) high-dimensional Gaussian approximation to our estimate, and 3) then proving the validity of the pivotal and nonparametric bootstraps building on 1) and 2). Each of these steps relies on modern empirical process theory and high-dimensional Gaussian approximation techniques recently developed by Chernozhukov et al. (2014, 2017a). In particular, the pivotal bootstrap differs from the nonparametric or multiplier bootstraps that have been analyzed in the literature in the high-dimensional setup (Belloni et al., 2019a; Chernozhukov et al., 2016b; Deng and Zhang, 2017; Chen and Kato, 2020), and proving the validity of the pivotal bootstrap requires a substantial work. Further, we employ a new multiplier inequality for the empirical process in Han and Wellner (2019) to establish the validity of the nonparametric bootstrap.

In summary, the present paper contributes to the literature on modal regression in twofold. First, we propose a new quantile-based conditional mode estimate that enjoys both desirable computational and statistical guarantees. The proposed estimator only requires solving a linear quantile regression problem and a one-dimensional optimization both of which can be solved efficiently. Second, we establish the theoretical validity of two bootstrap methods for a broad spectrum of inference tasks in a unified way. In particular, we propose a new resampling method (pivotal bootstrap) that builds on an insight into the specific structure of our estimate.

### 2.1.2 Literature review

Starting from the pioneering work of Sager and Thisted (1982), there is now a large literature on modal regression. There are two major approaches to estimating the conditional mode comparable to our method; one is linear modal regression where the conditional mode is assumed to be linear in covariates (Lee, 1989, 1993; Kemp and Santos-Silva, 2012; Yao and Li, 2014), and the other is nonparametric estimation (Yao et al., 2012; Chen et al., 2016; Yao and Xiang, 2016; Feng et al., 2020); see also Lee and Kim (1998); Manski (1991); Einbeck and Tutz (2006); Sasaki et al. (2016); Ho et al. (2017); Khardani and Yao (2017); Krief (2017) for alternative methods including semiparametric and Bayesian estimation. Lee (1989, 1993) assume symmetry of the error distribution to derive limit theorems for their proposed estimators, but the symmetry assumption implies that the conditional mean, median, and mode coincide, thereby significantly reducing the complexity of estimating the conditional mode. Kemp and Santos-Silva (2012) and Yao and Li (2014) consider an alternative estimator defined by minimizing a kernel-based loss function for linear modal regression and develop limit distribution theory for the estimator without assuming symmetry of the error distribution. However, the optimization problem of Kemp and Santos-Silva (2012) and Yao and Li (2014) is (multidimensional and) nonconvex, and while they propose EM-type algorithms to compute their estimators, "there is no guarantee that the algorithm will converge to the global optimal solution" (Yao and Li, 2014, p. 659). Compared with the method of Kemp and Santos-Silva (2012) and Yao and Li (2014), all three methods (including ours) enjoy the same rate of convergence, while our method is computationally attractive since linear quantile regression can be formulated as a linear programming problem (Koenker, 2005), and minimizing the estimated derivative of the

conditional quantile function is a one-dimensional optimization problem both of which can be solved accurately and efficiently.

Yao et al. (2012) consider local linear estimation of the conditional mode but their Condition (A6) is essentially the symmetry assumption on the error distribution, which makes their problem statistically equivalent to conditional mean estimation. Chen et al. (2016) study nonparametric estimation of the conditional mode based on kernel density estimation (KDE), and develop nonparametric bootstrap inference for their KDE-based estimate. The nonparametric estimation is able to avoid model misspecification. Chen et al. (2016) also allow for multiple local modes, while we assume the existence of the unique global mode at each design point of interest. Thus, the setup of Chen et al. (2016) is more general than ours. However, the convergence rate of the KDE-based estimate of Chen et al. (2016) is slow even when the dimension of the covariate vector is moderately large ("curse of dimensionality"). Specifically, the convergence rate of the Chen et al. (2016) estimate is at best $n^{-2/(p+7)}$ where $p$ is the number of continuous covariates under the assumption of four times differentiability of the conditional density, while our estimate can achieve the $n^{-2/7}$ rate (up to logarithmic factors when evaluated under the uniform norm) assuming three times differentiability of the conditional density (albeit assuming a linear quantile regression model). Finally, Chen et al. (2016) also consider the application of the nonparametric bootstrap to inference on the conditional mode. However, our estimator is substantially different from their estimator and requires different analysis to establish the validity of the nonparametric bootstrap.

The present paper builds on (but substantially differs from) the recent work of Ota et al. (2019), which proposes a different quantile-based estimate of the

conditional mode and develops pointwise limit distribution theory for their estimator. Contrary to ours, Ota et al. (2019) directly use the linear quantile regression estimate and minimize its difference quotient (as the linear quantile regression estimate is not smooth in the quantile index), which makes a substantial difference between their asymptotic analysis and ours. Indeed, Ota et al. (2019) show that the rate of convergence of their estimate is at best $n^{-1/4}$ that is slower than our $n^{-2/7}$ rate, and find that the pointwise limit distribution is a scale transformation of nonstandard Chernoff's distribution. The nonstandard limit distribution poses a substantial challenge in inference using their estimate and Ota et al. (2019) only consider pointwise inference using a general purpose subsampling method (Politis et al., 1999). We overcome this limitation by employing kernel smoothing, and further, develop a model-based bootstrap method (pivotal bootstrap) that enables us to deal with much broader inference tasks including simultaneous confidence intervals and significance testing.

This paper also builds on and contributes to the quantile regression literature. Quantile regression provides a comparatively full picture of how the covariates impact the conditional distribution of a response variable and has wide applications (Koenker, 2017). In particular, the pivotal bootstrap of the present paper is related to Parzen et al. (1994); Chernozhukov et al. (2009); He (2017); Belloni et al. (2019a) who study resampling-based inference methods that build on (conditionally) pivotal influence functions in the quantile regression setup. Their scopes and methods are, however, substantially different from ours. To the best of our knowledge, exploiting pivotal influence functions to make inference for modal regression is new.

## 2.2 Mode estimation via smoothed quantile regression

We begin with the setup and define our estimator. We are interested in making inference on the conditional mode of a scalar response variable $Y \in \mathbb{R}$ given a $d$-dimensional covariate vector $\boldsymbol{X} \in \mathbb{R}^d$. We will initially assume that the dimension $d$ is fixed in Section 2.3, but consider the extension to the case with $d = d_n \to \infty$ in Section 2.5. In what follows, we assume that there exists a conditional density of $Y$ given $\boldsymbol{X}$, $f(y \mid \boldsymbol{x})$, which is (at least) continuous in $y$ for each design point $\boldsymbol{x}$. We are interested in making inference on the conditional mode over a compact subset $\mathcal{X}_0$ of the support of $\boldsymbol{X}$. We assume that for each $\boldsymbol{x} \in \mathcal{X}_0$, there exists a unique global mode $m(\boldsymbol{x})$, i.e., $m(\boldsymbol{x})$ is the unique maximizer of the function $y \mapsto f(y \mid \boldsymbol{x})$,

$$m(\boldsymbol{x}) = \operatorname*{arg\,max}_{y \in \mathbb{R}} f(y \mid \boldsymbol{x}). \tag{2.1}$$

Our strategy to estimate the conditional mode is based on "inverting" a quantile regression model. For $\tau \in (0, 1)$, let $Q_{\boldsymbol{x}}(\tau)$ denote the conditional $\tau$-quantile of $Y$ given $\boldsymbol{X}$. Observe that the derivative of the conditional quantile function with respect to the quantile index $\tau$ coincides with the reciprocal of the conditional density at $Q_{\boldsymbol{x}}(\tau)$, i.e.,

$$s_{\boldsymbol{x}}(\tau) := Q'_{\boldsymbol{x}}(\tau) := \frac{\partial Q_{\boldsymbol{x}}(\tau)}{\partial \tau} = \frac{1}{f(Q_{\boldsymbol{x}}(\tau) \mid \boldsymbol{x})}. \tag{2.2}$$

This suggests that the conditional mode $m(\boldsymbol{x})$ can be obtained by minimizing the "sparsity" function $s_{\boldsymbol{x}}(\tau) := Q'_{\boldsymbol{x}}(\tau)$. Specifically, let $\tau_{\boldsymbol{x}}$ denote the minimizer of $s_{\boldsymbol{x}}(\cdot)$, i.e.,

$$\tau_{\boldsymbol{x}} = \operatorname*{arg\,min}_{\tau \in (0,1)} s_{\boldsymbol{x}}(\tau).$$

Then, we arrive at the expression $m(\boldsymbol{x}) = Q_{\boldsymbol{x}}(\tau_{\boldsymbol{x}})$. Hence, estimation of $m(\boldsymbol{x})$ reduces to estimation of $Q_{\boldsymbol{x}}(\cdot)$ and $\tau_{\boldsymbol{x}}$.

To estimate the conditional quantile function, we assume a linear quantile model, i.e.,

$$Q_{\boldsymbol{x}}(\tau) = \boldsymbol{x}^T \beta(\tau), \quad \tau \in (0,1).$$

Suppose that we are given i.i.d. observations $(Y_1, \boldsymbol{X}_1), \ldots, (Y_n, \boldsymbol{X}_n)$ of $(Y, \boldsymbol{X})$. We estimate the slope vector $\beta(\tau)$ by the standard quantile regression estimator (Koenker and Bassett, 1978),

$$\hat{\beta}(\tau) = \underset{\beta \in \mathbb{R}^d}{\arg\min} \sum_{i=1}^{n} \rho_\tau(Y_i - \boldsymbol{X}_i^T \beta), \tag{2.3}$$

where $\rho_\tau(u) = u\{\tau - I(u \leq 0)\}$ is the check function. However, the plug-in estimator $\check{Q}_{\boldsymbol{x}}(\tau) := \boldsymbol{x}^T \hat{\beta}(\tau)$ for the conditional quantile function is not smooth in $\tau$. To overcome this difficulty, we propose to smooth the naive estimator $\check{Q}_{\boldsymbol{x}}(\tau)$ by a kernel function, and estimate $\tau_{\boldsymbol{x}}$ by minimizing the derivative of the smoothed quantile estimator. To this end, let $K : \mathbb{R} \to \mathbb{R}$ be a kernel function (a function that integrates to 1) that is smooth and supported in $[-1, 1]$ (see Assumption 1 (vii) in the following for more details). For a given sequence of bandwidth parameters $h = h_n \to 0$, we modify the naive estimator $\check{Q}_{\boldsymbol{x}}(\tau)$ by

$$\hat{Q}_{\boldsymbol{x}}(\tau) := \int_{\tau-h}^{\tau+h} \check{Q}_{\boldsymbol{x}}(t) K_h(\tau - t) dt, \ \tau \in [\epsilon, 1-\epsilon],$$

where $K_h(\cdot) := h^{-1} K(\cdot/h)$ and $\epsilon \in (0, 1/2)$ is some small user-chosen parameter. The restriction of the range of $\tau$ is to avoid the boundary problem. Since $K$ is supported in $[-1, 1]$, the integral $\int_{\tau-h}^{\tau+h}$ above can be formally replaced by $\int_{\mathbb{R}}$ with the convention that $\check{Q}_{\boldsymbol{x}}(t) = 0$ for $t \notin (0, 1)$.

Then, we can estimate $s_{\boldsymbol{x}}(\tau)$ by differentiating $\hat{Q}_{\boldsymbol{x}}(\tau)$, $\hat{s}_{\boldsymbol{x}}(\tau) := \hat{Q}'_{\boldsymbol{x}}(\tau)$, and

estimate $\tau_{\boldsymbol{x}}$ by minimizing $\hat{s}_{\boldsymbol{x}}(\tau)$,

$$\hat{\tau}_{\boldsymbol{x}} := \underset{\tau \in [\epsilon, 1-\epsilon]}{\arg\min}\, \hat{s}_{\boldsymbol{x}}(\tau).$$

By the smoothness of $K(\cdot)$, the map $\tau \mapsto \hat{s}_{\boldsymbol{x}}(\tau)$ is smooth, so $\hat{\tau}_{\boldsymbol{x}}$ is guaranteed to exist by compactness of $[\epsilon, 1-\epsilon]$. Finally, we propose to estimate the conditional mode $m(\boldsymbol{x})$ by a plug-in method:

$$\hat{m}(\boldsymbol{x}) := \hat{Q}_{\boldsymbol{x}}\left(\hat{\tau}_{\boldsymbol{x}}\right).$$

Some remarks on the proposed estimator are in order.

**Remark 1** (Linear quantile regression)**.** The linear quantile regression model is common in the quantile regression literature and can cover many data generating processes (see Remark 1 in Ota et al. 2019). Importantly, the linear quantile regression problem can be solved efficiently since the optimization problem (2.3) can be formulated as a (parametric) linear programming problem whose solution path can be computed efficiently even for large-scale datasets (Koenker, 2005). Having said that, the linear specification of the conditional quantile function is not essential and the theoretical results developed in the following Section 2.3 and Section 2.5 can be extended to nonlinear quantile regression models.

**Remark 2** (Comparison with other estimators)**.** Compared with linear modal regression, our setting allows for nonlinear conditional mode functions even though the conditional quantile function is assumed linear in $\boldsymbol{x}$ (see Remark 1 in Ota et al. (2019)). In fact, under linear quantile assumption, $m(\boldsymbol{x}) = \boldsymbol{x}^T\beta(\tau_{\boldsymbol{x}})$ and $\beta(\tau_{\boldsymbol{x}})$ is allowed to be a (possibly nonlinear) function of $\boldsymbol{x}$. In addition, computation of linear modal regression involves non-convex optimization (Yao and Li, 2014; Cheng, 1995; Einbeck and Tutz, 2006), while the proposed method only relies on linear quantile regression that can be formulated as a linear programming problem, and an one-dimensional optimization. Chen et al. (2016) show

the convergence rate $O_P(h^2 + n^{-1/2}h^{-(p+3)/2})$ for the KDE-based mode estimator, where $h$ is the KDE bandwidth parameter and $p$ is the number of continuous covariates. This implies slow convergence for even moderate dimensions which is the price of a more nonparametric approach. In contrast, we show that the convergence rate of our estimator is $O_P(h^2 + n^{-1/2}h^{-3/2})$ for any fixed dimension $d$ and thus our estimator is free from the "curse of dimensionality".

## 2.3 Main results

### 2.3.1 Notation and conditions

We use $U(0,1)$ and $N(\mu, \Sigma)$ to denote the uniform distribution on $(0,1)$ and the normal distribution with mean $\mu$ and covariance matrix $\Sigma$, respectively. We use $\|\cdot\|$, $\|\cdot\|_1$, $\|\cdot\|_\infty$ to denote the Euclidean, $\ell^1$, and $\ell^\infty$-norms, respectively. For a smooth function $f(x)$, we write $f^{(r)}(x) = \partial^r f(x)/\partial x^r$ for any integer $r \geq 0$ with $f^{(0)} = f$. For vectors $a = (a_1, \ldots, a_L)^T, b = (b_1, \ldots, b_L)^T \in \mathbb{R}^L$, we write $a \leq b$ if $a_\ell \leq b_\ell$ for all $1 \leq \ell \leq L$.

Let $\mathcal{X} \subset \mathbb{R}^d$ denote the support of $\boldsymbol{X}$ and let $\mathcal{X}_0 \subset \mathcal{X}$ be the set over which we make inference on the conditional mode. In this section the dimension $d$ of $\boldsymbol{X}$ is assumed to be fixed. Recall the baseline assumption in the last section that we are given i.i.d. observations $(Y_1, \boldsymbol{X}_1), \ldots, (Y_n, \boldsymbol{X}_n)$ of $(Y, \boldsymbol{X})$ where the conditional distribution of $Y$ given $\boldsymbol{X}$ has a unique mode and satisfies the linear quantile regression model. We make the following additional assumption.

**Assumption 1.** *(i) The set $\mathcal{X}_0$ is compact in $\mathbb{R}^d$; (ii) For any $\boldsymbol{x} \in \mathcal{X}_0$, $\tau_{\boldsymbol{x}} \in (\epsilon, 1-\epsilon)$; (iii) The covariate vector $\boldsymbol{X}$ has finite $q$-th moment, $\mathbb{E}[\|\boldsymbol{X}\|^q] < \infty$, for some $q \in [4, \infty)$,*

*and the Gram matrix $\mathbb{E}[\boldsymbol{X}\boldsymbol{X}^T]$ is positive definite; (iv) The conditional density $f(y \mid \boldsymbol{x})$ is three times continuously differentiable with respect to $y$ for each $\boldsymbol{x} \in \mathcal{X}$. Let $f^{(j)}(y \mid \boldsymbol{x}) = \partial^j f(y \mid \boldsymbol{x})/\partial y^j$ for $j = 0, 1, 2, 3$. There exits a constant $C_1$ such that $|f^{(j)}(y \mid \boldsymbol{x})| \leq C_1$ for all $j = 0, 1, 2, 3$ and $(y, \boldsymbol{x}) \in \mathbb{R} \times \mathcal{X}$; (v) There exists a positive constant $c_1$ (that may depend on $\epsilon$) such that $f(y \mid \boldsymbol{x}) \geq c_1$ for all $y \in [Q_{\boldsymbol{x}}(\epsilon/2), Q_{\boldsymbol{x}}(1-\epsilon/2)]$ and $\boldsymbol{x} \in \mathcal{X}$; (vi) There exists a positive constant $c_2$ such that $-f^{(2)}(m(\boldsymbol{x}) \mid \boldsymbol{x}) \geq c_2$ for all $\boldsymbol{x} \in \mathcal{X}_0$; (vii) The kernel function $K$ is three times differentiable, symmetric, and supported in $[-1, 1]$; (viii) The bandwidth $h = h_n \to 0$ satisfies that $nh^5/\log n \to \infty$.*

Condition (i) is innocuous (recall that $\mathcal{X}_0$ is not the support of $\boldsymbol{X}$). Condition (ii) excludes the extreme quantile case where $\tau_{\boldsymbol{x}} \to 0$ or $1$ for some sequence of $\boldsymbol{x}$. Condition (iii) is a moment condition on the covariate vector $\boldsymbol{X}$. Conditions (iv) and (v) are standard smoothness conditions on the conditional density $f(\cdot \mid \boldsymbol{x})$ in the quantile regression literature (Koenker, 2005). Similar conditions appear in Chen et al. (2016) and Ota et al. (2019). Smoothness of $f(\cdot \mid \boldsymbol{x})$ implies smoothness of conditional quantile function $Q_{\boldsymbol{x}}(\tau)$. Indeed, under Conditions (iv) and (v), $Q_{\boldsymbol{x}}(\tau)$ is four-times continuously differentiable. Condition (vi) ensures that the conditional mode $m(\boldsymbol{x})$ as a solution to the optimization problem (2.1) is nondegenerate. Condition (vi) also ensures that the map $\boldsymbol{x} \mapsto s''_{\boldsymbol{x}}(\tau_{\boldsymbol{x}})$ is bounded away from zero on $\mathcal{X}_0$, as

$$s''_{\boldsymbol{x}}(\tau) = Q^{(3)}_{\boldsymbol{x}}(\tau) = \frac{3f^{(1)}(Q_{\boldsymbol{x}}(\tau) \mid \boldsymbol{x}) - f(Q_{\boldsymbol{x}}(\tau) \mid \boldsymbol{x}) f^{(2)}(Q_{\boldsymbol{x}}(\tau) \mid \boldsymbol{x})}{f(Q_{\boldsymbol{x}}(\tau) \mid \boldsymbol{x})^5}$$

and $f^{(1)}(Q_{\boldsymbol{x}}(\tau_{\boldsymbol{x}}) \mid \boldsymbol{x}) = f^{(1)}(m(\boldsymbol{x}) \mid \boldsymbol{x}) = 0$. It is important to note that we only require Condition (vi) to hold for $\boldsymbol{x} \in \mathcal{X}_0$, the set of design points we make inference on. A similar condition to Condition (vi) also appears in Chen et al. (2016). Conditions (vii) and (viii) are concerned with the kernel function $K$ and the bandwidth $h_n$. We will use the biweight kernel $K(t) = \frac{15}{16}(1-t^2)^2 I(|t| <$

1) in our numerical studies. Condition (viii) ensures $\hat{Q}^{(3)}_{\boldsymbol{x}}(\tau)$ to be (uniformly) consistent; see Lemma 6 in Appendix.

### 2.3.2 Uniform asymptotic linear representation

In this section, we derive a uniform asymptotic linear representation for our estimator $\hat{m}(\boldsymbol{x})$, which will be a building block for the pivotal bootstrap. Define

$$J(\tau) := \mathbb{E}[f\left(\boldsymbol{X}^T\beta(\tau) \mid \boldsymbol{X}\right)\boldsymbol{X}\boldsymbol{X}^T].$$

By Assumption 1 (iii) and (v), the minimum eigenvalue of the matrix $J(\tau)$ is bounded away from zero for $\tau \in [\epsilon, 1-\epsilon]$. Further, for $(u, \boldsymbol{x}') \in (0,1) \times \mathbb{R}^d$, define

$$\psi_{\boldsymbol{x}}(u, \boldsymbol{x}') := -\frac{s_{\boldsymbol{x}}(\tau_{\boldsymbol{x}})}{s''_{\boldsymbol{x}}(\tau_{\boldsymbol{x}})\sqrt{h}} K'\left(\frac{\tau_{\boldsymbol{x}} - u}{h}\right) \boldsymbol{x}^T J(\tau_{\boldsymbol{x}})^{-1}\boldsymbol{x}',$$

which will serve as an influence function for our estimator $\hat{m}(\boldsymbol{x})$. Let $\kappa = \int t^2 K(t)dt$.

**Proposition 1** (Uniform asymptotic linear representation)**.** *Under Assumption 1, the following asymptotic linear representation holds uniformly in $\boldsymbol{x} \in \mathcal{X}_0$:*

$$\begin{aligned}
&\hat{m}(\boldsymbol{x}) - m(\boldsymbol{x}) + \frac{s_{\boldsymbol{x}}(\tau_{\boldsymbol{x}})s^{(3)}_{\boldsymbol{x}}(\tau_{\boldsymbol{x}})}{2s''_{\boldsymbol{x}}(\tau_{\boldsymbol{x}})}\kappa h^2 + o_P(h^2) \\
&\quad = \frac{1}{nh^{3/2}}\sum_{i=1}^{n} \psi_{\boldsymbol{x}}(U_i, \boldsymbol{X}_i) + O_P(n^{-1/2}h^{-1} + n^{-1}h^{-4}\log n),
\end{aligned}$$

*where $U_1, \ldots, U_n \sim U(0,1)$ i.i.d. independent of $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$. In addition, we have*

$$\sup_{\boldsymbol{x}\in\mathcal{X}_0}\left|\frac{1}{nh^{3/2}}\sum_{i=1}^{n}\psi_{\boldsymbol{x}}(U_i, \boldsymbol{X}_i)\right| = O_P(n^{-1/2}h^{-3/2}\sqrt{\log n}).$$

The influence function $\psi_{\boldsymbol{x}}(U_i, \boldsymbol{X}_i)$ has mean zero when $h \le \min\{\tau_{\boldsymbol{x}}, 1-\tau_{\boldsymbol{x}}\}$ which holds for sufficiently large $n$, since

$$\int_0^1 K'\left(\frac{\tau_{\boldsymbol{x}} - u}{h}\right)du = h\int_{(\tau_{\boldsymbol{x}}-1)/h}^{\tau_{\boldsymbol{x}}/h} K'(u)du = h\int_{\mathbb{R}} K'(u)du = 0 \tag{2.4}$$

and by independence between $U_i$ and $\boldsymbol{X}_i$. Proposition 1 in particular implies pointwise asymptotic normality of the proposed estimator.

**Corollary 1** (Pointwise asymptotic normality)**.** *Suppose that Assumption 1 holds. Then, for any fixed $\boldsymbol{x} \in \mathcal{X}_0$, we have*

$$\sqrt{nh^3}\left[\hat{m}(\boldsymbol{x}) - m(\boldsymbol{x}) + \frac{s_{\boldsymbol{x}}(\tau_{\boldsymbol{x}})s^{(3)}_{\boldsymbol{x}}(\tau_{\boldsymbol{x}})}{2s''_{\boldsymbol{x}}(\tau_{\boldsymbol{x}})}\kappa h^2 + o_P(h^2)\right] \xrightarrow{d} N(0, V_{\boldsymbol{x}}),$$

*where $V_{\boldsymbol{x}} = s_{\boldsymbol{x}}(\tau_{\boldsymbol{x}})^2\mathbb{E}[(\boldsymbol{x}^T J(\tau_{\boldsymbol{x}})^{-1}\boldsymbol{X})^2]\kappa_1/s''_{\boldsymbol{x}}(\tau_{\boldsymbol{x}})^2$ and $\kappa_1 = \int K'(t)^2 dt$.*

Proposition 1 shows that the uniform convergence rate of the proposed estimator is $O_P(n^{-1/2}h^{-3/2}\sqrt{\log n} + h^2)$, which is dimension-free (i.e., independent of $d$). If we choose $h \sim (n/\log n)^{-1/7}$, which balances between $n^{-1/2}h^{-3/2}\sqrt{\log n}$ and $h^2$, then the rate reduces to $O_P((n/\log n)^{-2/7})$.

### 2.3.3 Bootstrap inference

We consider simultaneous inference for the conditional mode at several design points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_L \in \mathcal{X}_0$, where $L$ is allowed to depend on $n$, i.e., $L = L_n \to \infty$. Indeed, we aim at developing a general inference framework to construct confidence sets for linear combinations of the vector $(m(\boldsymbol{x}_\ell))_{\ell=1}^L$. Specifically, we consider making inference on $D(m(\boldsymbol{x}_\ell))_{\ell=1}^L$ where $D$ is a deterministic $M \times L$ matrix and the number of rows $M$ is also allowed to increase with $n$, i.e., $M = M_n \to \infty$. The following are a few examples of the matrix $D$. See also Examples 5 and 6 ahead for more details.

**Example 3** (Simultaneous confidence intervals)**.** Suppose that we are interested in constructing simultaneous confidence intervals for the conditional mode at

design points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_L$. Construction of such simultaneous confidence intervals requires approximating the distribution of the vector $(\hat{m}(\boldsymbol{x}_\ell) - m(\boldsymbol{x}_\ell))_{\ell=1}^L$, and thus $D = I_L$ ($L \times L$ identity matrix).

Another application is constructing simultaneous confidence intervals for partial effects of certain covariates on the conditional mode, i.e., the change of the conditional mode due to the change of one particular covariate while the rest of the covariates are controlled. Inference on partial effects is an important topic in econometrics and social science (Williams, 2012). For example, suppose that we have covariate $\boldsymbol{X} = (X_1, X_{-1})$ where $X_{-1}$ contains covariates other than $X_1$. Consider to construct simultaneous confidence intervals for partial effects of $X_1$ at $M$ different design points $x_1^{(1)}, \ldots, x_1^{(M)}$: $m(x_1^{(k)} + \delta, x_{-1}) - m(x_1^{(k)}, x_{-1})$ $(1 \leq k \leq M)$ for some small user-chosen $\delta$ and fixed $x_{-1}$. To this end, we need to approximate the distribution of $(\hat{m}(x_1^{(k)} + \delta, x_{-1}) - \hat{m}(x_1^{(k)}, x_{-1}))_{k=1}^M$. If we take $\boldsymbol{x}_{2k-1} = (x_1^{(k)} + \delta, x_{-1})$ and $\boldsymbol{x}_{2k} = (x_1^{(k)}, x_{-1})$ for $k = 1, \ldots M$, then the corresponding $D$ matrix is $D_c$ in (2.5).

**Example 4** (Testing significance of covariates)**.** Suppose first that we are interested in testing whether the conditional mode is constant over designs points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_L$, i.e., $m(\boldsymbol{x}_1) = \cdots = m(\boldsymbol{x}_L)$, which is equivalent to test $m(\boldsymbol{x}_{\ell+1}) - m(\boldsymbol{x}_\ell) = 0$ simultaneously for all $1 \leq \ell \leq L-1$ (this corresponds to testing lack of significance of all covariates). Calibrating critical values for such tests reduces to approximating the null distribution of the vector $(\hat{m}(\boldsymbol{x}_{\ell+1}) - \hat{m}(\boldsymbol{x}_\ell))_{\ell=1}^{L-1}$, and thus the matrix $D$ is $D_t$ in (2.5).

We can also consider testing significance of certain covariates on the conditional mode. For instance, suppose that we have three covariates (including 1): $\boldsymbol{X} = (1, X_1, X_2)^T$ with binary $X_2$ (i.e., $X_2 \in \{0, 1\}$), and we are interested

in testing lack of significance of the covariate $X_2$, i.e., $m(X_1, 0) = m(X_1, 1)$ (the constant 1 is omitted from the expression of $m(\boldsymbol{X})$). This can be carried out by picking designs points $x_1^{(1)}, \ldots, x_1^{(M)}$ from the support of $X_1$, and testing the simultaneous hypothesis that $m(x_1^{(k)}, 0) = m(x_1^{(k)}, 1)$ (or equivalently $m(x_1^{(k)}, 0) - m(x_1^{(k)}, 1) = 0$) for all $k = 1, \ldots, M$. Calibrating critical values for such tests requires us to approximate the distribution of $(\hat{m}(x_1^{(k)}, 0) - \hat{m}(x_1^{(k)}, 1))_{k=1}^M$. If we define $\boldsymbol{x}_{2k-1} = (x_1^{(k)}, 0)$ and $\boldsymbol{x}_{2k} = (x_1^{(k)}, 1)$ for $k = 1, \ldots, M$, then the corresponding $D$ matrix is the same as $D_c$ in (2.5).

$$D_c = \underbrace{\begin{pmatrix} 1 & -1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & -1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \cdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 & -1 \end{pmatrix}}_{M \times 2M}; \qquad D_t = \underbrace{\begin{pmatrix} 1 & -1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & -1 & \cdots & 0 & 0 \\ \vdots & & \ddots & \cdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -1 \end{pmatrix}}_{(L-1) \times L}. \tag{2.5}$$

To cover above applications in a unified way, we consider to approximate the distribution of $D(\hat{m}(\boldsymbol{x}_\ell) - m(\boldsymbol{x}_\ell))_{\ell=1}^L$. We will first show that, under regularity conditions, $\sqrt{nh^3}D(\hat{m}(\boldsymbol{x}_\ell) - m(\boldsymbol{x}_\ell))_{\ell=1}^L$ can be approximated by an $L$-dimensional Gaussian vector uniformly over the hyperrectangles in $\mathbb{R}^L$, even when $L$ and $M$ are possibly much larger then $n$. This approximating Gaussian distribution is infeasible in practice since its covariance matrix is unknown. To deal with this difficulty, we propose to further approximate the sampling distribution by a novel pivotal bootstrap or the conventional nonparametric bootstrap.

**Gaussian approximation**

Define $\Psi_i := (\psi_{\boldsymbol{x}_1}(U_i, \boldsymbol{X}_i), \ldots, \psi_{\boldsymbol{x}_L}(U_i, \boldsymbol{X}_i))^T$ and $\Sigma := \mathbb{E}[\Psi_i \Psi_i^T]$. For $k = 1, \ldots, M$, let $D_k^T$ denote the $k$-th row of the matrix $D$. We may assume without loss of generality that each row $D_k$ is nonzero. Further, we will assume that the matrix $D$ is sparse in the sense that the number of nonzero elements of each row $D_k$ is of constant order, which is satisfied in all the examples discussed above. We are primarily interested in inference for the vector $((m(\boldsymbol{x}_\ell))_{\ell=1}^L$, so we normalize the coordinates of the vector by their approximate standard deviations (technically the normalization does not matter for the Gaussian approximation, but we will replace the approximate standard deviations by their estimates in the bootstrap, whose effect has to be taken care of). Let $S_k := \{\ell \in \{1, \ldots, L\} : D_{k,\ell} \neq 0\}$ denote the support of $D_k$. Define the normalization matrix $\Gamma := \mathrm{diag}\{\Gamma_1, \ldots, \Gamma_M\}$ and set $A = (A_1, \ldots, A_M)^T := \Gamma^{-1} D$. In particular, if we take $\Gamma_k = \sqrt{D_k^T \Sigma D_k}$ for $k = 1, \ldots, M$, which corresponds to the standard deviation of $D_k^T \Psi_i$, such choice of $A$ will result in a studentized statistic, while taking $\Gamma = I_M$ gives a non-studentized statistic.

Related to the matrix $D$ and $\Gamma$, we make the following assumption.

**Assumption 2.** *(i)* $\max_{1 \leq k \leq M} |S_k| = O(1)$ *and* $\max_{1 \leq k \leq M; 1 \leq \ell \leq L} |D_{k,\ell}| = O(1)$*; (ii) There exists a fixed constant* $c_3 > 0$ *such that* $\min_{1 \leq k \leq M} D_k^T \Sigma D_k \geq c_3$*; (iii) There exists a fixed constant* $c_4 > 0$ *such that* $c_4 \leq \min_{1 \leq k \leq M} \Gamma_k \leq \max_{1 \leq k \leq M} \Gamma_k = O(1)$.

Condition (i) is a sparsity assumption on the matrix $D$ discussed above. The conditions Condition (ii) excludes the situation where $D_k^T \Psi_i$ has vanishing variance. Condition (iii) imposes a mild condition on the normalization matrix $\Gamma$ which is automatically satisfied for both studentized and non-studentized cases

under the previous two conditions.

The following theorem derives a Gaussian approximation result.

**Theorem 1** (Gaussian approximation)**.** *Suppose that Assumptions 1 and 2 hold. In addition, assume that*

$$\frac{\log^7(Mn)}{nh} \bigvee \frac{\log^3(Mn)}{n^{1-2/q}h} \bigvee \frac{(\log^2 n)\log M}{nh^5} \to 0 \quad \textit{and} \quad (nh^7 \vee h)\log M \to 0. \tag{2.6}$$

*Then, we have*

$$\sup_{b\in\mathbb{R}^M} \left| \mathbb{P}\left( A\sqrt{nh^3}(\hat{m}(\boldsymbol{x}_\ell) - m(\boldsymbol{x}_\ell))_{\ell=1}^L \le b \right) - \mathbb{P}\left(AG \le b\right) \right| \to 0,$$

*where $G$ is an $L$-dimensional Gaussian random vector with mean $0$ and covariance $\Sigma$.*

Condition (2.6) allows $M$ to be much larger than $n$, i.e., $M \gg n$. The condition that $nh^7 \log M \to 0$ is an "undersmoothing" condition that ensures that the deterministic bias is negligible relative to the stochastic error. This condition can be relaxed by assuming additional smoothness conditions on the conditional density and using higher order kernels. We do not pursue this extension for brevity. Discussion on the bandwidth selection can be found in Section 2.4.1.

The proof of Theorem 1 can be found in the Appendix. The proof builds on the uniform asymptotic linear representation developed in Proposition 1 coupled with the high dimensional Gaussian approximation techniques developed in Chernozhukov et al. (2014, 2017a). From Theorem 1, we see that the distribution of $A\sqrt{nh^3}(\hat{m}(\boldsymbol{x}_\ell) - m(\boldsymbol{x}_\ell))_{\ell=1}^L$ can be approximated by the distribution of $AG$ uniformly over the rectangles. Still, the distribution of $AG$ is unknown since the covariance matrix of $G$ is unknown. We will use a new bootstrap called the pivotal bootstrap or nonparametric bootstrap to further estimate the distribution of $AG$.

**Remark 3** (Limit distribution of maximum deviation)**.** It is of interest to find a limit distribution of the maximum deviation, $\zeta_n := \max_{1\le\ell\le L}\sqrt{nh^3}|\hat{m}(\boldsymbol{x}_\ell) - m(\boldsymbol{x}_\ell)|/\sigma_{\boldsymbol{x}_\ell}$ with $\sigma^2_{\boldsymbol{x}} = \mathbb{E}[\psi_{\boldsymbol{x}}(U,\boldsymbol{X})^2]$, when $L = L_n \to \infty$ after a suitable normalization. Such a limit distribution enables us to find analytical critical values for simultaneous confidence intervals. Indeed, combining Theorem 1 with extreme value theory (cf. Leadbetter et al., 1983), we can derive a limit distribution for the maximal deviation under additional regularity conditions, cf. Proposition 5 in Appendix A and discussion there.

**Remark 4** (Conditioning on $\boldsymbol{X}_i$'s)**.** Inspection of the proof of Theorem 1 shows that a version of the conclusion of Theorem 1 continues to hold conditionally on the covariate vectors $\boldsymbol{X}_1,\dots,\boldsymbol{X}_n$, with minor modifications to the regularity conditions:

$$\sup_{b\in\mathbb{R}^M}\left|\mathbb{P}\left(A\sqrt{nh^3}(\hat{m}(\boldsymbol{x}_\ell) - m(\boldsymbol{x}_\ell))_{\ell=1}^L \le b \mid \boldsymbol{X}_1,\dots,\boldsymbol{X}_n\right) - \mathbb{P}\left(AG\le b\right)\right| \xrightarrow{P} 0. \tag{2.7}$$

Thus, combined with the consistency of the pivotal and nonparametric bootstraps, the size and coverage guarantees of inference methods constructed from those bootstraps continue to hold conditionally on the covariate vectors $\boldsymbol{X}_1,\dots,\boldsymbol{X}_n$. The proof of the result (2.7) is indeed similar to the validity of the pivotal bootstrap (see Theorem 2 below), as the pivotal bootstrap is essentially using the randomness of $U_1,\dots,U_n$ alone. We omit the details for brevity.

**Pivotal bootstrap**

The proof of Theorem 1 shows that the distribution of $G$ comes from approximating the distribution of the process

$$\boldsymbol{x} \mapsto \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_{\boldsymbol{x}}(U_i, \boldsymbol{X}_i) \tag{2.8}$$

at $\boldsymbol{x} \in \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_L\}$. Importantly, the process (2.8) is "pivotal" in the sense that its distribution is completely known up to some estimable nuisance parameters given $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ since $U_1, \ldots, U_n$ are independent $U(0,1)$ random variables. The baseline idea of the pivotal bootstrap is to simulate the pivotal process (2.8) (given the data) to estimate the distribution of $G$ by generating $U(0,1)$ random variables.

To implement the pivotal bootstrap, we first have to estimate the nuisance parameters. We consider to estimate the matrix $J(\tau) = \mathbb{E}[f(\boldsymbol{X}^T\beta(\tau) \mid \boldsymbol{X})\boldsymbol{X}\boldsymbol{X}^T]$ by Powell's kernel method (Powell, 1986), i.e., $\hat{J}(\tau) := n^{-1}\sum_{i=1}^{n} \check{K}_{\check{h}_n}(Y_i - \boldsymbol{X}_i^T\hat{\beta}(\tau))\boldsymbol{X}_i\boldsymbol{X}_i^T$, where $\check{K} : \mathbb{R} \to \mathbb{R}$ is a kernel function and $\check{h}_n$ is a bandwidth. For simplicity of exposition, we will use $\check{K} = K$ and $\check{h}_n = h$. Then, we shall estimate the influence function $\psi_{\boldsymbol{x}}$ by

$$\hat{\psi}_{\boldsymbol{x}}(u, \boldsymbol{x}') := -\frac{\hat{s}_{\boldsymbol{x}}(\hat{\tau}_{\boldsymbol{x}})}{\hat{s}''_{\boldsymbol{x}}(\hat{\tau}_{\boldsymbol{x}})\sqrt{h}} K'\left(\frac{\hat{\tau}_{\boldsymbol{x}} - u}{h}\right) \boldsymbol{x}^T \hat{J}(\hat{\tau}_{\boldsymbol{x}})^{-1}\boldsymbol{x}',$$

where $\hat{s}''_{\boldsymbol{x}}(\tau)$ is the second derivative of $\hat{s}_{\boldsymbol{x}}(\tau)$ with respect to $\tau$.

The pivotal bootstrap reads as follows. Generate $U_1, \ldots, U_n \sim U(0,1)$ i.i.d. that are independent of the data $\mathcal{D}_n := (Y_i, \boldsymbol{X}_i)_{i=1}^n$. We denote the conditional probability $\mathbb{P}(\cdot \mid \mathcal{D}_n)$ and conditional expectation $\mathbb{E}[\cdot \mid \mathcal{D}_n]$ by $\mathbb{P}_{|\mathcal{D}_n}(\cdot)$ and $\mathbb{E}_{|\mathcal{D}_n}[\cdot]$, respectively. Define

$$\hat{\Psi}_i := \left(\hat{\psi}_{\boldsymbol{x}_1}(U_i, \boldsymbol{X}_i), \ldots, \hat{\psi}_{\boldsymbol{x}_L}(U_i, \boldsymbol{X}_i)\right)^T \quad \text{and} \quad \hat{\Sigma} := \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{|\mathcal{D}_n}[\hat{\Psi}_i\hat{\Psi}_i^T].$$

Then, we shall estimate the distribution of $AG$ (or $n^{-1/2}\sum_{i=1}^n A\Psi_i$) by the conditional distribution of $n^{-1/2}\sum_{i=1}^n \hat{A}\hat{\Psi}_i$ given the data $\mathcal{D}_n$, where $\hat{A} = \hat{\Gamma}^{-1}D$ and $\hat{\Gamma} = \mathrm{diag}\{\hat{\Gamma}_1, \ldots, \hat{\Gamma}_M\}$ is some estimator of $\Gamma$ (for example, equation (2.9)) that achieves sufficiently fast convergence rate (see Theorem 2 for details). The conditional distribution can be simulated with arbitrary precision. The following theorem establishes consistency of the pivotal bootstrap over the rectangles.

**Theorem 2** (Validity of pivotal bootstrap)**.** *Suppose that Assumptions 1 and 2 hold with $q > 4$ in Condition (v) in Assumption 1. In addition, assume that*

*(i).* $\max_{1\le k\le M} |\hat{\Gamma}_k - \Gamma_k| = O_P(n^{-1/2}h^{-5/2}\sqrt{\log n} + h)$.

*(ii).* $\frac{\log^7(Mn)}{n^{1-2/q}h} \bigvee \frac{\log^3(Mn)}{n^{1-4/q}h} \bigvee \frac{(\log n)\log^4 M}{nh^5} \to 0 \quad \textit{and} \quad h\log^2 M \to 0.$

*Then, we have*

$$\sup_{b\in\mathbb{R}^M} \left| \mathbb{P}_{|\mathcal{D}_n}\left(n^{-1/2}\textstyle\sum_{i=1}^n \hat{A}\hat{\Psi}_i \le b\right) - \mathbb{P}\left(AG \le b\right) \right| \xrightarrow{P} 0.$$

The proof of Theorem 2 can be found in Appendix. The proof of Theorem 2 is nontrivial and does not follow directly from existing results since the pivotal bootstrap differs from the nonparametric or multiplier bootstraps that have been analyzed in the literature in the high-dimensional setup. The proof consists of two steps. First, noting that $\hat{\Psi}_1, \ldots, \hat{\Psi}_n$ are independent with mean zero conditionally on the data $\mathcal{D}_n$ (cf. equation (2.4)), we apply the high dimensional CLT conditionally on $\mathcal{D}_n$ to approximate the conditional distribution of $n^{-1/2}\sum_{i=1}^n \hat{A}\hat{\Psi}_i$ by the conditional Gaussian distribution $N(0, \hat{A}\hat{\Sigma}\hat{A}^T)$. Second, we compare the $N(0, \hat{A}\hat{\Sigma}\hat{A}^T)$ distribution with $AG \sim N(0, A\Sigma A^T)$ by a Gaussian comparison technique.

**Remark 5** (Choice of $\hat{A}$)**.** For the non-studentized case, i.e., $\Gamma = I_M$, we can simply take $\hat{A} = D$. For the studentized case, i.e., $\Gamma_k = \sqrt{D_k^T \Sigma D_k}$ $(1 \leq k \leq M)$, we can estimate $\Gamma$ by

$$\hat{\Gamma} := \mathrm{diag}\left\{\sqrt{D_1^T \hat{\Sigma} D_1}, \ldots, \sqrt{D_M^T \hat{\Sigma} D_M}\right\}, \tag{2.9}$$

and compute $\hat{A}$ accordingly. We can show that this $\hat{\Gamma}$ satisfies Condition (i) of Theorem 2 (cf. Lemma 12 in Appendix). In practice, $\hat{\Sigma}$ can be approximated by simulating uniform random variables and then $\hat{\Gamma}$ can be computed according to (2.9).

As a byproduct of the proof of Theorem 2, we can show that the conclusion of Theorem 1 continues to hold even if the matrix $A$ acting on $(\hat{m}(\boldsymbol{x}_\ell) - m(\boldsymbol{x}_\ell))_{\ell=1}^L$ is replaced by its estimate $\hat{A}$.

**Proposition 2.** *Suppose that the conditions of Theorem 1 together with Condition (i) in the statement of Theorem 2 hold. In addition, assume that*

$$\frac{(\log n) \log^2 M}{nh^5} \to 0 \quad \textit{and} \quad h \log M \to 0.$$

*Then, we have*

$$\sup_{b \in \mathbb{R}^M} \left| \mathbb{P}\left( \hat{A}\sqrt{nh^3}(\hat{m}(\boldsymbol{x}_\ell) - m(\boldsymbol{x}_\ell))_{\ell=1}^L \leq b \right) - \mathbb{P}\left(AG \leq b\right) \right| \to 0.$$

In what follows, we discuss applications of the pivotal bootstrap to constructions of pointwise and simultaneous confidence intervals and testing using studentized statistics.

**Example 5** (Simultaneous confidence intervals)**.** Consider construction of a simultaneous confidence interval for $m(\boldsymbol{x}_1), \ldots, m(\boldsymbol{x}_L)$. In this case, $D = I_L$ $(M = L)$, $A = \mathrm{diag}\{1/\sigma_{\boldsymbol{x}_1}, \ldots, 1/\sigma_{\boldsymbol{x}_L}\}$, and $\hat{A} = \mathrm{diag}\{1/\hat{\sigma}_{\boldsymbol{x}_1}, \ldots, 1/\hat{\sigma}_{\boldsymbol{x}_L}\}$, where

$\sigma_{\boldsymbol{x}}^2 = \mathbb{E}[\psi_{\boldsymbol{x}}(U, \boldsymbol{X})^2]$ and $\hat{\sigma}_{\boldsymbol{x}}^2 = n^{-1}\sum_{i=1}^n \mathbb{E}_{|\mathcal{D}_n}[\hat{\psi}_{\boldsymbol{x}}(U_i, \boldsymbol{X}_i)^2]$. Then, Proposition 2 and Theorem 2 imply that, for $G = (g_1, \ldots, g_L)^T \sim N(0, \Sigma)$,

$$\sup_{b\in\mathbb{R}} \left| \mathbb{P}\left( \max_{1\le\ell\le L} \left| \sqrt{nh^3}(\hat{m}(\boldsymbol{x}_\ell) - m(\boldsymbol{x}_\ell))/\hat{\sigma}_{\boldsymbol{x}_\ell} \right| \le b \right) - \mathbb{P}\left( \max_{1\le\ell\le L} |g_\ell/\sigma_{\boldsymbol{x}_\ell}| \le b \right) \right| \to 0, \text{ and}$$
$$\sup_{b\in\mathbb{R}} \left| \mathbb{P}_{|\mathcal{D}_n}\left( \max_{1\le\ell\le L} \left| n^{-1/2}\textstyle\sum_{i=1}^n \hat{\psi}_{\boldsymbol{x}_\ell}(U_i, \boldsymbol{X}_i)/\hat{\sigma}_{\boldsymbol{x}_\ell} \right| \le b \right) - \mathbb{P}\left( \max_{1\le\ell\le L} |g_\ell/\sigma_{\boldsymbol{x}_\ell}| \le b \right) \right| \xrightarrow{P} 0. \tag{2.10}$$

Denoting by

$$\hat{q}_{1-\alpha} = \text{conditional } (1-\alpha)\text{-quantile of } \max_{1\le\ell\le L} \left| n^{-1/2}\textstyle\sum_{i=1}^n \hat{\psi}_{\boldsymbol{x}_\ell}(U_i, \boldsymbol{X}_i)/\hat{\sigma}_{\boldsymbol{x}_\ell} \right|,$$

we can show that the data-dependent rectangle (interval when $L = 1$)

$$\prod_{\ell=1}^L \left[ \hat{m}(\boldsymbol{x}_\ell) \pm \frac{\hat{\sigma}_{\boldsymbol{x}_\ell}}{\sqrt{nh^3}} \hat{q}_{1-\alpha} \right]$$

contains the vector $(m(\boldsymbol{x}_\ell))_{\ell=1}^L$ with probability approaching $1-\alpha$.

Formally, the coverage guarantee of the preceding confidence rectangle follows from

$$\mathbb{P}\left( \max_{1\le\ell\le L} \left| \sqrt{nh^3}(\hat{m}(\boldsymbol{x}_\ell) - m(\boldsymbol{x}_\ell))/\hat{\sigma}_{\boldsymbol{x}_\ell} \right| \le \hat{q}_{1-\alpha} \right) \to 1-\alpha. \tag{2.11}$$

The latter (2.11) follows from the preceding convergence result (2.10) coupled with Lemma 1 in Appendix (note: since in general $\max_{1\le\ell\le L} |g_\ell/\sigma_{\boldsymbol{x}_\ell}|$ need not have a limit distribution, it is not immediate that the former (2.10) implies the latter (2.11); cf. Lemma 23.3 in van der Vaart (2000)). A similar analysis can be done for constructing simultaneous confidence intervals for partial effects of certain covariates.

**Example 6** (Testing significance of covariates)**.** Consider testing the hypothesis $H_0 : m(\boldsymbol{x}_1) = \cdots = m(\boldsymbol{x}_L)$ for some $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_L \in \mathcal{X}_0$. In this case, the matrix $D$ is given by $D_t$ in (2.5) with $M = L-1$, and $A(\hat{m}(\boldsymbol{x}_\ell) - m(\boldsymbol{x}_\ell))_{\ell=1}^L = \big((\hat{m}(\boldsymbol{x}_{\ell+1}) -$

$\hat{m}(\boldsymbol{x}_\ell))/\sigma_{\boldsymbol{x}_{\ell+1},\boldsymbol{x}_\ell}\big)_{\ell=1}^{L-1}$ under $H_0$, where $\sigma^2_{\boldsymbol{x}_{\ell+1},\boldsymbol{x}_\ell} = \mathbb{E}[(\psi_{\boldsymbol{x}_{\ell+1}} - \psi_{\boldsymbol{x}_\ell})^2(U, \boldsymbol{X})]$. Let $\hat{\sigma}^2_{\boldsymbol{x}_{\ell+1},\boldsymbol{x}_\ell} = n^{-1}\sum_{i=1}^n \mathbb{E}_{|\mathcal{D}_n}[(\hat{\psi}_{\boldsymbol{x}_{\ell+1}} - \hat{\psi}_{\boldsymbol{x}_\ell})^2(U_i, \boldsymbol{X}_i)]$. We shall consider the test of the form

$$\max_{1\le\ell\le L-1} \frac{\sqrt{nh^3}|\hat{m}(\boldsymbol{x}_{\ell+1}) - \hat{m}(\boldsymbol{x}_\ell)|}{\hat{\sigma}_{\boldsymbol{x}_{\ell+1},\boldsymbol{x}_\ell}} > c \;\Rightarrow\; \text{reject } H_0 \tag{2.12}$$

for some critical value $c$. To calibrate critical values, we may use the pivotal bootstrap. For a given level $\alpha \in (0,1)$, let

$$\hat{c}_{1-\alpha} = \text{conditional } (1-\alpha)\text{-quantile of } \max_{1\le\ell\le L-1} \left| n^{-1/2}\textstyle\sum_{i=1}^n (\hat{\psi}_{\boldsymbol{x}_{\ell+1}} - \hat{\psi}_{\boldsymbol{x}_\ell})(U_i, \boldsymbol{X}_i)/\hat{\sigma}_{\boldsymbol{x}_{\ell+1},\boldsymbol{x}_\ell} \right|.$$

Then, Proposition 2 and Theorem 2 guarantee that, under regularity conditions, the test (2.12) with $c = \hat{c}_{1-\alpha}$ has level approaching $\alpha$ if $H_0$ is true (cf. the discussion at the end of the preceding example). The case where the $D$ matrix is given by $D_c$ in (2.5) is similar; we omit the details for brevity.

**Nonparametric bootstrap**

In this section, we consider and analyze the nonparametric (empirical) bootstrap to approximate the sampling distribution of our estimator or the approximating Gaussian distribution $AG$ that appears in Theorem 1. The nonparametric bootstrap proceeds as follows. We draw $n$ i.i.d. bootstrap samples $(Y_1^*, \boldsymbol{X}_1^*), \ldots, (Y_n^*, \boldsymbol{X}_n^*)$ from the empirical distribution of $(Y_1, \boldsymbol{X}_1), \ldots, (Y_n, \boldsymbol{X}_n)$. For a design point $\boldsymbol{x} \in \mathcal{X}_0$, we denote the mode estimator computed from the bootstrap samples by $\hat{m}^*(\boldsymbol{x})$. Then, we shall estimate the distribution of $AG$ by the conditional distribution of $\hat{A}\sqrt{nh^3}(\hat{m}^*(\boldsymbol{x}_\ell) - \hat{m}(\boldsymbol{x}_\ell))_{\ell=1}^L$ given the data $\mathcal{D}_n$, where we define the same $\hat{A}$ as in the pivotal bootstrap. The following theorem establishes consistency of the nonparametric bootstrap over the rectangles.

**Theorem 3** (Validity of nonparametric bootstrap)**.** *Suppose that Assumptions 1*

*and 2 hold with $q > 4$ in Condition (v) in Assumption 1. In addition, assume that, for arbitrarily small $\gamma > 0$,*

*(i).* $\max_{1\le k\le M} |\hat{\Gamma}_k - \Gamma_k| = O_P(n^{-1/2}h^{-5/2}\sqrt{\log n} + h)$.

*(ii).* $\frac{\log^7(Mn)}{n^{1-2/q}h} \bigvee \frac{\log^3(Mn)}{n^{1-4/q}h} \bigvee \frac{\log^4 M}{n^{1-\gamma}h^5} \to 0$ *and* $(h\log M \vee nh^7)\log M \to 0$.

*Then, we have*

$$\sup_{b\in\mathbb{R}^M} \left| \mathbb{P}_{|\mathcal{D}_n} \left( \hat{A}\sqrt{nh^3}(\hat{m}^*(\boldsymbol{x}_\ell) - \hat{m}(\boldsymbol{x}_\ell))_{\ell=1}^L \le b \right) - \mathbb{P}\left(AG \le b\right) \right| \xrightarrow{P} 0.$$

The proof of Theorem 3 can be found in Appendix. The proof consists of the following steps. First, we establish a uniform linear representation for $(\hat{m}^*(\boldsymbol{x}_\ell) - \hat{m}(\boldsymbol{x}_\ell))_{\ell=1}^L$ based on a Bahadur representation for the nonparametric bootstrap quantile regression estimator (see Lemma 13); Then we follow a similar proof strategy to Theorem 1 while conditioning on the data $\mathcal{D}_n$ that requires a more involved analysis than Theorem 1.

**Remark 6** (Comparison with the pivotal bootstrap)**.** The consistency of two bootstrap methods are established under fairly similar conditions. However, the nonparametric bootstrap can be computationally more demanding since it requires computing mode estimates on sufficiently many bootstrap samples. In contrast, the pivotal bootstrap only requires estimating nuisance parameters once and evaluating the influence functions repeatedly by generating uniform random variables, which can be easily parallelized and adapted to the distributed setting. Therefore, the pivotal bootstrap can be computationally more attractive than the nonparametric bootstrap. Our simulation results also demonstrate the computational advantage of the pivotal bootstrap over the nonparametric bootstrap, cf. Appendix B.5.1.

## 2.4 Numerical examples

### 2.4.1 Simulation results

In this section, we present the numerical performance of the pivotal bootstrap using synthetic data. Due to the space limitation, we defer the simulation results for the nonparametric bootstrap and pivotal bootstrap testing (Example 6) to Appendices E.1 and E.3 in the supplementary material. We start with discussing implementation details, in particular the bandwidth selection.

#### Implementation details

In our simulation study, we use the biweight kernel, $K(t) = \frac{15}{16}(1-t^2)^2 I(|t| < 1)$, and use $\epsilon = 0.1$ when computing our modal estimator. We estimate the matrix $J(\tau)$ by $\hat{J}(\tau) = (2n\check{h})^{-1} \sum_{i=1}^{n} I(|Y_i - \boldsymbol{X}_i^T \hat{\beta}(\tau)| \leq \check{h}) \boldsymbol{X}_i \boldsymbol{X}_i^T$, where $\check{h}$ is set to be the default bandwidth in *quantreg* package in R (the theory does not require the kernel used to estimate $J(\tau)$ to be smooth). For the minimization of the sparsity function, we used the R function *optimize()* with the computed derivative of the smoothed quantile function as the input. We find that computing $\hat{s}''(\tau_{\boldsymbol{x}})$ by differentiating $\hat{Q}_{\boldsymbol{x}}(\tau)$ three times tends to be unstable in the finite sample. Instead, we use the alternative expression $s''(\tau_{\boldsymbol{x}}) = -f^{(2)}(Q_{\boldsymbol{x}}(\tau_{\boldsymbol{x}}) \mid \boldsymbol{x}) s_{\boldsymbol{x}}(\tau_{\boldsymbol{x}})^4$ and estimate the derivative $f^{(2)}(\cdot \mid \boldsymbol{x})$ by a kernel method as in Remark 9 of Ota et al. (2019) (we plug in $\hat{Q}_{\boldsymbol{x}}(\hat{\tau}_{\boldsymbol{x}})$ and $\hat{s}_{\boldsymbol{x}}(\hat{\tau}_{\boldsymbol{x}})$ for $Q_{\boldsymbol{x}}(\tau_{\boldsymbol{x}})$ and $s_{\boldsymbol{x}}(\tau_{\boldsymbol{x}})$, respectively). We defer more implementation details of nuisance parameter estimation to Appendix G.

Finally, we discuss bandwidth selection. Corollary 1 implies that the approx-

imate MSE of $\hat{m}(\boldsymbol{x})$ is

$$\left[\frac{s_{\boldsymbol{x}}(\tau_{\boldsymbol{x}})s_{\boldsymbol{x}}^{(3)}(\tau_{\boldsymbol{x}})}{2s_{\boldsymbol{x}}''(\tau_{\boldsymbol{x}})}\kappa h^2\right]^2 + \frac{\kappa_1 s_{\boldsymbol{x}}(\tau_{\boldsymbol{x}})^2\mathbb{E}[(x^T J(\tau_{\boldsymbol{x}})^{-1}\boldsymbol{X})^2]}{nh^3 s_{\boldsymbol{x}}''(\tau_{\boldsymbol{x}})^2}.$$

The optimal $h$ that minimizes the above approximate MSE is given by

$$h_{\text{opt}}(\boldsymbol{x}) := \left[\frac{3\kappa_1 \boldsymbol{x}^T J(\tau_{\boldsymbol{x}})^{-1}\mathbb{E}[\boldsymbol{X}\boldsymbol{X}^T]J(\tau_{\boldsymbol{x}})^{-1}\boldsymbol{x}}{\kappa^2 s_{\boldsymbol{x}}^{(3)}(\tau_{\boldsymbol{x}})^2}\right]^{1/7} n^{-1/7}.$$

Here we make some remarks on the optimal bandwidth. First, we note that direct use of $h_{\text{opt}}$ will result in an asymptotic bias and a bias-correction will be needed. However, the asymptotic bias contains high order derivatives of the conditional quantile function that are hard to estimate. Hence, we recommend a smaller bandwidth to be used in the finite sample implementation. In our numerical analysis, we multiply $h_{\text{opt}}$ by $0.8$ to correct for too large bandwidths. We start with an initial bandwidth $h_{\text{ini}} = 0.8 \times n^{-1/7}$ to get the initial estimate $\hat{\tau}_{\boldsymbol{x}}^0$ for $\tau_{\boldsymbol{x}}$ and replace $J(\tau_{\boldsymbol{x}})$ and $\mathbb{E}[\boldsymbol{X}\boldsymbol{X}^T]$ in $h_{\text{opt}}$ with $\hat{J}(\hat{\tau}_{\boldsymbol{x}}^0)$ and $n^{-1}\sum_{i=1}^n \boldsymbol{X}_i\boldsymbol{X}_i^T$ respectively. Considering that estimation of the fourth derivative of the conditional quantile function is highly unstable, we adopt a "rule of thumb" method by using the fourth derivative of the quantile function of the standard normal distribution, i.e., plugging in $(\Phi^{-1}(\tau))^{(4)}$ for $s_{\boldsymbol{x}}^{(3)}(\tau)$ regardless of different design points, where $\Phi$ is the distribution function of $N(0,1)$. This will lead to an estimate of $h_{\text{opt}}$. We iterate the process one more time to construct the final computed bandwidth. For the simultaneous inference on multiple design points, we take the bandwidth to be the median of the pointwise bandwidths at those design points. Our empirical results show that the above bandwidth selection approach works reasonably well.

**Pointwise confidence intervals**

We will consider two different models which correspond to linear and nonlinear mode functions respectively. Suppose that $Y$ and $\boldsymbol{X} = (1, X_1)$ are generated from either of the following models,

- (Linear modal function) $Y = 1 + 3X_1 + \sigma(X_1)\xi$,
- (Nonlinear modal function) $Y = 3U^3 - 3X_1U^2 + 3X_1U$.

In the linear modal function case, we take $\sigma(x) = 1 + 2x$. For the distribution of $\xi$, we consider two cases: $\xi \sim N(0, 1)$ (*lmNormal* model) and $\log(\xi) \sim N(1, 0.64)$ (*lmLognormal* model). These two cases are of interest since the conditional mode coincides with the conditional mean in the first case while they are different in the second. In particular, $m(\boldsymbol{X}) = 1 + 3X_1$ for the *lmNormal* model and $m(\boldsymbol{X}) = 1+3X_1+(1+2X_1)e^{0.36}$ for the *lmLognormal* model, both of which are linear in $\boldsymbol{X}$. Similar models are considered in the simulation analyses of Yao and Li (2014) and Ota et al. (2019). For the nonlinear modal function case (*Nonlinear* model), we take $U \sim U(0, 1)$ and thus $m(\boldsymbol{X}) = -2X_1^3/9 + X_1^2$, which is nonlinear in $\boldsymbol{X}$. We generate the covariate $X_1 \sim U(0, 1)$ in both linear modal models and $X_1 \sim U(0, 3)$ for the nonlinear modal model.

For each model, we construct $95\%$ and $99\%$ confidence intervals for the conditional mode. For the *lmNormal* and *lmLognormal* models, we consider the following design points $\boldsymbol{x} = (1, 0.3)$, $(1, 0.5)$ and $(1, 0.7)$, while for the *Nonlinear* model, we consider $\boldsymbol{x} = (1, 0.7)$, $(1, 0.9)$ and $(1, 1.1)$. We consider different sample sizes ranging from 500 to 2000 and repeat computing the confidence intervals under different sample sizes for $500$ times. The resulting empirical coverage probabilities and interval length statistics are reported in Tables 2.1 to 2.3. In

the simulation, we find that some of the computed confidence intervals are extremely large, especially when the sample size is comparatively small ($n = 500$) due to the unstable estimation of high order derivatives of the conditional quantile function. Therefore, we report the median length of the confidence intervals to exclude the influence of those extreme results. We also present the interquartile range of the lengths of the computed confidence intervals.

Table 2.1: Simulation results for pointwise confidence intervals for *lmNormal* model. For each case, the results are computed based on 500 simulated datasets with 500 bootstrap iterations. IQR represents "interquartile range".

| Design point | Sample size | Coverage probability | | Median length | | IQR | |
|---|---|---|---|---|---|---|---|
| | | 95% | 99% | 95% | 99% | 95% | 99% |
| $X_1$=0.3 | $n = 500$ | 94.4% | 97.6% | 1.17 | 1.55 | 1.23 | 1.62 |
| | $n = 1000$ | 96% | 98.2% | 0.94 | 1.23 | 0.92 | 1.19 |
| | $n = 2000$ | 96% | 98.2% | 0.79 | 1.03 | 0.71 | 0.93 |
| $X_1$=0.5 | $n = 500$ | 95.2% | 98% | 1.43 | 1.87 | 1.29 | 1.64 |
| | $n = 1000$ | 95.6% | 98.8% | 1.12 | 1.47 | 0.99 | 1.3 |
| | $n = 2000$ | 96% | 98.6% | 0.92 | 1.18 | 0.69 | 0.92 |
| $X_1$=0.7 | $n = 500$ | 90.4% | 94.2% | 1.42 | 1.89 | 1.67 | 2.21 |
| | $n = 1000$ | 93% | 96.4% | 1.29 | 1.67 | 1.70 | 2.21 |
| | $n = 2000$ | 93.4% | 96.6% | 1.02 | 1.35 | 1.08 | 1.45 |

Table 2.2: Simulation results for pointwise confidence intervals for *lmLognormal* model. For each case, the results are computed based on 500 simulated datasets with 500 bootstrap iterations. IQR represents "interquartile range".

| Design point | Sample size | Coverage probability | | Median length | | IQR | |
|---|---|---|---|---|---|---|---|
| | | 95% | 99% | 95% | 99% | 95% | 99% |
| $X_1$=0.3 | $n = 500$ | 92.2% | 96.6% | 3.00 | 3.73 | 4.05 | 4.82 |
| | $n = 1000$ | 94.4% | 97.4% | 2.12 | 2.77 | 1.70 | 2.09 |
| | $n = 2000$ | 92.4% | 96.4% | 2.03 | 2.65 | 1.16 | 1.49 |
| $X_1$=0.5 | $n = 500$ | 93.8% | 96.8% | 3.70 | 4.65 | 4.61 | 5.69 |
| | $n = 1000$ | 90.8% | 96.2% | 2.45 | 3.26 | 2.07 | 2.76 |
| | $n = 2000$ | 96.4% | 98.8% | 2.00 | 2.62 | 1.17 | 1.46 |
| $X_1$=0.7 | $n = 500$ | 90.2% | 94.6% | 4.58 | 5.63 | 5.60 | 6.39 |
| | $n = 1000$ | 92% | 95.8% | 2.89 | 3.68 | 2.19 | 2.82 |
| | $n = 2000$ | 93.6% | 97.4% | 2.33 | 3.03 | 1.64 | 2.01 |

From Tables 2.1 to 2.3, the bootstrap confidence intervals achieve satisfying coverage probabilities in all three scenarios. We point out that, in each case, the

Table 2.3: Simulation results for pointwise confidence intervals for *Nonlinear* model. For each case, the results are computed based on 500 simulated datasets with 500 bootstrap iterations. IQR represents "interquartile range".

| Design point | Sample size | Coverage probability | | Median length | | IQR | |
|---|---|---|---|---|---|---|---|
| | | 95% | 99% | 95% | 99% | 95% | 99% |
| $X_1$=0.7 | $n = 500$ | 95.2% | 97.2% | 0.80 | 1.01 | 0.97 | 1.12 |
| | $n = 1000$ | 95% | 97.6% | 0.66 | 0.84 | 0.77 | 0.93 |
| | $n = 2000$ | 94.8% | 97% | 0.50 | 0.64 | 0.55 | 0.62 |
| $X_1$=0.9 | $n = 500$ | 91.6% | 95.2% | 0.75 | 0.95 | 1.03 | 1.25 |
| | $n = 1000$ | 91.4% | 95.8% | 0.64 | 0.83 | 0.85 | 1.07 |
| | $n = 2000$ | 95.2% | 97.8% | 0.58 | 0.75 | 0.80 | 1.04 |
| $X_1$=1.1 | $n = 500$ | 92% | 95.6% | 0.87 | 1.14 | 1.18 | 1.47 |
| | $n = 1000$ | 93% | 96.6% | 0.69 | 0.89 | 0.92 | 1.15 |
| | $n = 2000$ | 94.4% | 96.8% | 0.57 | 0.74 | 0.79 | 1.00 |

coverage probabilities at $X_1 = 0.7$ are slightly lower than the other two design points under the same sample size. This is because large $X_1$ results in a large variance of $Y$ which makes the estimation more difficult. We report the mean squared error of our conditional mode estimator, $\hat{m}_{\boldsymbol{x}}$ in Appendix B.5.2 to verify this. However, the pivotal bootstrap still achieves approximately nominal coverage probabilities in such situations when the sample size is sufficiently large. Besides, we note that the length of the confidence intervals and its variability decrease with the growing sample size for each design point across all three scenarios, which agrees with our asymptotic theories. We also report oracle pivotal bootstrap confidence intervals in Appendix B.5.4 where we plug in $s''_{\boldsymbol{x}}(\tau_{\boldsymbol{x}})$ using the underlying true density or conditional quantile function. From the results there, we can see a decrease in the length and interquartile range (in particular, the latter) for the oracle confidence intervals comparing with the results presented above under the same setting. Therefore, we conclude that the estimation of the nuisance parameters may impact the performance of the confidence intervals significantly.

**Approximate confidence band**

In this section, we investigate the finite sample performance of the pivotal bootstrap in simultaneous inference problems. In particular, we construct approximate confidence bands for the three different models considered in Section 2.4.1. To build an approximate confidence band, we compute simultaneous confidence intervals over a equally spaced grid of $X_1$. Specifically, we consider a grid with 21 points over interval $[0.4, 0.6]$ for the *lmNormal* and *lmLognormal* models and over interval $[0.6, 1.2]$ for the *Nonlinear* model.

We repeat the simulation 500 times for each model and calculate the empirical coverage probabilities and the median lengths defined by taking the median of the median length of the simultaneous confidence intervals in one simulation. The median is used to reduce the influence of potential extreme results in the simulations. The resulting empirical coverage probabilities and median lengths of the approximate confidence bands for each model are presented in Table 2.4.

Table 2.4: Simulation results for approximate confidence bands for *lmNormal*, *lmLognormal* and *Nonlinear* models. For each case, the results are computed based on 500 simulated datasets with 500 bootstrap iterations.

| Models | Sample size | Coverage probability | | Median length | |
|---|---|---|---|---|---|
| | | 95% | 99% | 95% | 99% |
| lmNormal | $n = 500$ | 93.4% | 97% | 1.71 | 2.16 |
| | $n = 1000$ | 94.6% | 97.8% | 1.35 | 1.70 |
| | $n = 2000$ | 94.8% | 98.4% | 1.07 | 1.36 |
| lmLognormal | $n = 500$ | 95.2% | 97.8% | 6.30 | 7.77 |
| | $n = 1000$ | 94.6% | 98% | 4.53 | 5.72 |
| | $n = 2000$ | 97.4% | 99.2% | 3.48 | 4.35 |
| Nonlinear | $n = 500$ | 96.6% | 98.6% | 1.69 | 2.01 |
| | $n = 1000$ | 95.6% | 98.2% | 1.13 | 1.36 |
| | $n = 2000$ | 96.6% | 99.2% | 0.84 | 1.02 |

From Table 2.4, the approximate confidence bands are able to capture the modes simultaneously with probability close to the nominal probability for

large sample sizes. Additionally, similar to the pointwise confidence interval, the lengths of the confidence bands decrease while the sample size grows.

### 2.4.2 U.S. wage data

In this section, we apply the pivotal bootstrap inference framework on a real US wage data. The data are extracted from U.S. 1980 1% metro sample from the Integrated Public Use Microdata Series (IPUMS) website (Ruggles et al., 2020) and the dataset used in our analysis is provided in the supplemental material. We defer more details of the extracted dataset to Appendix H. In the following, the response $Y$ is the real log annual wage (wage), and the regressor $\boldsymbol{X}$ consists of the highest grade of schooling (edu), age (age) and marital status (marital_status).

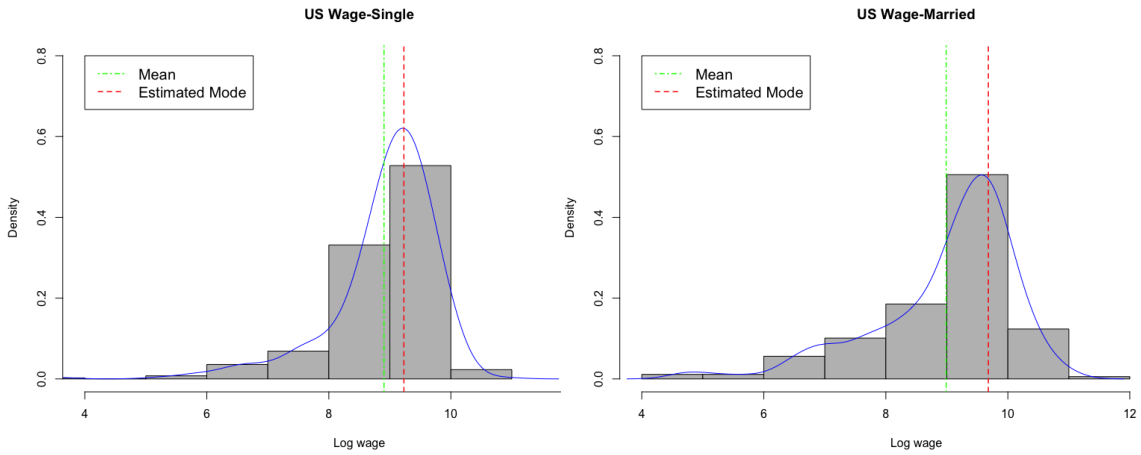
We investigate whether the most common wage given the same education and age is different in single and married people. Specifically, we take the two other covariates, education and age, to be the full-sample mode of each covariate and estimate the resulting conditional mode of these two groups. The estimation results are presented in Table 2.5.

Table 2.5: Mode estimates and the mode difference confidence intervals of the two groups.

| Estimated mode of wage | | Difference confidence intervals | |
|---|---|---|---|
| Single | Married | 95% | 99% |
| 9.23 | 9.68 | $(0.13, 0.79)$ | $(0.07, 0.84)$ |

To provide an intuitive evaluation of the estimation, in Figure 2.1, we collect the people with mode values of education and age from the two groups and plot KDE-based density estimates superimposed on histograms of their log annual

wage, respectively. The estimated modes (based on our estimator) and sample means are also highlighted in Figure 2.1.




(a) Single people (b) Married people

Figure 2.1: Histograms of log annual wage for single and married people with mode values of education and age based on U.S. 1980 1% metro sample data.

From Figure 2.1, we have several observations. First, both conditional distributions are skewed, and as argued in Kemp and Santos-Silva (2012), the mode would be a more intuitive measure of central tendency for such skewed data. Second, our modal estimator provides accurate estimations of conditional modes for both groups. We also present the confidence intervals for the difference of the modes of these two groups (the mode wage of single people minus the mode wage of married people) in Table 2.5. Since $0$ is not contained in both 95% and 99% confidence intervals, we conclude that the difference of the conditional modes between the two groups is statistically significant under those two nominal levels. Therefore, the marital status can possibly be a significant factor contributing to the mode of people's wage which can be of social interest worth further research.

## 2.5 Extension to the increasing dimension case

In this section, we extend the theoretical analysis to the case where the dimension $d$ of the covariate vector is allowed to increase with the sample size $n$, i.e., $d = d_n \to \infty$. Such situations arise when we approximate conditional quantile function $Q_{\boldsymbol{x}}(\tau)$ by a linear combination of basis functions and the approximation error is negligible (in fact, the theory of this section holds as long as the approximation error is at most of the order as the remainder term in the Bahadur representation; see Lemma 16 in Appendix). In this case, $\boldsymbol{X}$ is generated as basis functions of a fixed dimensional genuine covariate $\boldsymbol{Z}$, i.e., $\boldsymbol{X} = W(\boldsymbol{Z})$, where vector $W(\boldsymbol{Z})$ includes transformations of $\boldsymbol{Z}$ that have good approximation properties such as Fourier series, splines, and wavelets; cf. Belloni et al. (2015a, 2019a). It is then of interest to draw simultaneous confidence intervals for the conditional mode along with values of $\boldsymbol{Z}$ which has fixed dimension although the dimension of $\boldsymbol{X}$ increases with $n$.

We first modify Assumption 1 to accommodate the case where $d = d_n \to \infty$. In what follows, constants refer to nonrandom numbers independent of $n$.

**Assumption 3.** *(i) There exists a constant $C_2 \geq 1$ such that $C_2^{-1}\sqrt{d} \leq \|\boldsymbol{x}\| \leq C_2\sqrt{d}$ for all $\boldsymbol{x} \in \mathcal{X}_0$; (ii) There exists $\epsilon_1 \in (\epsilon, 1/2)$ such that $\tau_{\boldsymbol{x}} \in [\epsilon_1, 1-\epsilon_1]$ for all $\boldsymbol{x} \in \mathcal{X}_0$; (iii) There exists a positive constant $C_3$ such that $\mathbb{P}(\|\boldsymbol{X}\| \leq C_3\sqrt{d}) = 1$. The Gram matrix $\mathbb{E}[\boldsymbol{X}\boldsymbol{X}^T]$ is positive definite with smallest eigenvalue $\lambda_{\min} \geq c_{\min} > 0$ and largest eigenvalue $\lambda_{\max} \leq c_{\max} < \infty$ for some constants $c_{\min}$ and $c_{\max}$; (iv) Conditions (iv)–(vii) in Assumption 1 hold; (v) For any $\delta > 0$, there exists a positive constant $c_4$ (that may depend on $\delta$) such that $\inf_{\boldsymbol{x}\in\mathcal{X}_0} \inf_{\tau\in[\epsilon,1-\epsilon];|\tau-\tau_{\boldsymbol{x}}|\geq\delta}\{s_{\boldsymbol{x}}(\tau) - s_{\boldsymbol{x}}(\tau_{\boldsymbol{x}})\} \geq c_4$; (vi) $d^4 = o(n^{1-c_5})$ for some $c_5 \in (0,1)$.*

Condition (i) requires the design points of interest to be of the same order $\sqrt{d}$. We assume condition (i) to state the results in a concise way, but the $\sqrt{d}$ order can be relaxed as long as $\inf_{\boldsymbol{x}\in\mathcal{X}_0}\|\boldsymbol{x}\|$ and $\sup_{\boldsymbol{x}\in\mathcal{X}_0}\|\boldsymbol{x}\|$ are of the same order. The modified condition (ii) is assumed to avoid boundary problems of $\tau_{\boldsymbol{x}}$ when the dimension increases. We also assume that $\|\boldsymbol{X}\|$ is bounded by $C_3\sqrt{d}$ to avoid some technicalities. In particular, under series approximation framework, this assumption is satisfied when $\boldsymbol{X}$ is generated from basis functions such as Fourier series, B-splines and wavelet series; cf. Belloni et al. (2015a). The condition on the Gram matrix is satisfied under mild conditions on the distribution of the genuine covariate $\boldsymbol{Z}$ and basis functions; cf. Belloni et al. (2019a). Condition (v) is a global identification condition on $\tau_{\boldsymbol{x}}$ that is needed to verify the uniform consistency of $\hat{\tau}_{\boldsymbol{x}}$. If $d$ is fixed, then Condition (v) follows automatically as $\boldsymbol{x}\mapsto\tau_{\boldsymbol{x}}$ is continuous under Assumption 1 (see the proof of Lemma 8), but if $d=d_n\to\infty$, then $s_{\boldsymbol{x}}$ and $\tau_{\boldsymbol{x}}$ depend on $n$, so that we require Condition (v). Condition (vi) is used to guarantee the Bahadur representation of $\hat{\beta}(\tau)$; cf. Theorem 2 in Belloni et al. (2019a).

Redefine $\Psi_i$ as $\Psi_i := (\xi_{\boldsymbol{x}_1}(U_i,\boldsymbol{X}_i),\ldots,\xi_{\boldsymbol{x}_L}(U_i,\boldsymbol{X}_i))^T$ with

$$\xi_{\boldsymbol{x}}(u,\boldsymbol{x}') := \frac{s_{\boldsymbol{x}}(\tau_{\boldsymbol{x}})}{s''_{\boldsymbol{x}}(\tau_{\boldsymbol{x}})\sqrt{dh}}\int \boldsymbol{x}^T J(t)^{-1}\boldsymbol{x}'\{t-I(U\le t)\}K''\left(\frac{\tau_{\boldsymbol{x}}-t}{h}\right)dt.$$

Further, redefine the matrices $\Sigma$, $\Gamma$, and $A$ as in Section 2.3.3 corresponding to the new definition of $\Psi_i$. For simplicity, we focus here on the studentized case where $\Gamma_k = \sqrt{D_k^T\Sigma D_k}$ for $k=1,\ldots,M$. The reason to work with $\xi_{\boldsymbol{x}}$ instead of $\psi_{\boldsymbol{x}}$ is to better control the residual term in the proof of high dimensional Gaussian approximation result. Normalization by $\sqrt{d}$ ensures that the norm of $\boldsymbol{x}/\sqrt{d}$ is bounded on $\mathcal{X}_0$. The Gaussian approximation with $d=d_n\to\infty$ reads as follows.

**Theorem 4** (Gaussian approximation when $d = d_n \to \infty$)**.** *Suppose that Assumptions 2 and 3 hold and we also assume that*

$$\frac{d\log^7(Mn)}{nh} \bigvee \frac{d^4(\log^2 n)\log^2 M}{nh^2} \bigvee \frac{d^3(\log^2 n)\log M}{nh^5} \to 0 \quad \textit{and} \quad \frac{nh^7 \log M}{d} \to 0. \tag{2.13}$$

*Then, we have*

$$\sup_{b\in\mathbb{R}^M} \left| \mathbb{P}\left( A\sqrt{nh^3 d^{-1}}(\hat{m}(\boldsymbol{x}_\ell) - m(\boldsymbol{x}_\ell))_{\ell=1}^L \le b \right) - \mathbb{P}\left(AG \le b\right) \right| \to 0, \quad \textit{with } G \sim N(0,\Sigma).$$

Suppose that $\log M = O(\log n)$; then Condition (2.13) reduces to

$$\frac{d^4 \log^4 n}{nh^2} \bigvee \frac{d^3 \log^3 n}{nh^5} \to 0 \quad \text{and} \quad \frac{nh^7 \log n}{d} \to 0.$$

If we take $h = (n/d)^{-1/7}(\log n)^{-2}$, then the condition on $d$ reduces to $d^8 \cdot \text{polylog}(n) = o(n)$. As before, this condition can be relaxed by assuming additional smoothness conditions on the conditional density and using higher order kernels. Similar conditions on $d$ appear in the analysis of resampling methods for quantile regression under increasing dimensions; see, e.g., Theorem 5 in Belloni et al. (2019a), where $d = o(n^{1/10})$.

We now establish the validity of the pivotal bootstrap. The theory for the nonparametric bootstrap can be shown similarly but we omit the details due to the space limit. Redefine $\hat{\Psi}_i = (\hat{\psi}_{\boldsymbol{x}_1}(U_i, \boldsymbol{X}_i), \ldots, \hat{\psi}_{\boldsymbol{x}_L}(U_i, \boldsymbol{X}_i))^T$ with

$$\hat{\psi}_{\boldsymbol{x}}(u, \boldsymbol{x}') := -\frac{\hat{s}_{\boldsymbol{x}}(\hat{\tau}_{\boldsymbol{x}})}{\hat{s}''_{\boldsymbol{x}}(\hat{\tau}_{\boldsymbol{x}})\sqrt{dh}} K'\left(\frac{\hat{\tau}_{\boldsymbol{x}} - U_i}{h}\right) \boldsymbol{x}^T \hat{J}(\hat{\tau}_{\boldsymbol{x}})^{-1} \boldsymbol{X}_i$$

Let $\hat{\Gamma}$ be as in (2.9) corresponding to the new definition of $\hat{\psi}_{\boldsymbol{x}}$, and let $\hat{A} = \hat{\Gamma}^{-1} D$.

**Theorem 5** (Validity of pivotal bootstrap when $d = d_n \to \infty$)**.** *Suppose that Assumptions 2 and 3 hold and we also assume that*

$$\frac{d\log^7(Mn)}{nh} \bigvee \frac{d^2(d \vee h^{-2})(\log n)\log^4 M}{nh^3} \to 0 \quad \textit{and} \quad h\log^2 M \to 0.$$

*Then, we have*

$$\sup_{b\in\mathbb{R}^M}\left|\mathbb{P}_{|\mathcal{D}_n}\left(n^{-1/2}\textstyle\sum_{i=1}^n\hat{A}\hat{\Psi}_i\le b\right)-\mathbb{P}\left(AG\le b\right)\right|\xrightarrow{P}0.$$

**Remark 7.** The pivotal bootstrap above is the same as the one under the fixed dimension case as the extra normalization by $\sqrt{d}$ is canceled by the multiplication by $\hat{A}$ (we introduced normalization by $\sqrt{d}$ to facilitate the proof).

## 2.6 Summary

In this paper, we study a novel pivotal bootstrap and the nonparametric bootstrap for simultaneous inference on conditional modes based on a kernel-smoothed Koenker-Bassett quantile estimator. Our bootstrap inference framework allows for simultaneous inference on multiple linear functions of different conditional modes. We establish the validity of the bootstrap inference in both fixed dimension and increasing dimension settings. The numerical results provide strong support of our theoretical results. Several interesting extensions remain, including the extension to time series or longitudinal data. In such settings, we need to modify the bootstraps and develop new technical tools to deal with dependent data. These are beyond the scope of the current paper and left for future research.

CHAPTER 3

# DISTRIBUTED INFERENCE FOR HIGH-DIMENSIONAL LAD REGRESSION VIA MULTI-ROUND AGGREGATION

## 3.1 Introduction

### 3.1.1 Overview

Explosive growth of the size and dimensionality of modern datasets brings hope of new scientific discoveries but statistical challenges as well. High-dimensional data analysis where the dimension of the data far exceeds the sample size has been an active research area in the past two decades Bühlmann and Van De Geer (2011); Hastie et al. (2015); Belloni and Chernozhukov (2011); Belloni et al. (2012). Meanwhile, massive datasets usually cannot be fit in a single machine due to memory constraints and therefore have to be processed on multiple machines. This fueled another line of research on the distributed statistical analysis Jordan et al. (2019); Wang et al. (2017a); Lee et al. (2017); Chen et al. (2019); Volgushev et al. (2019) that has wide application, for instance, in large-scale sensor networks. We refer to the aforementioned literature for more motivating real-world examples.

Quantile regression, a principal methodology to study the conditional distribution, has wide applications in various research areas, including survival analysis Xu et al. (2017), epidemiology Wei et al. (2019) and economics Belloni et al. (2019a), due to its ability to accommodate heavy error tail conditions and account for the heteroskedasticity Volgushev et al. (2019). We consider estimation

and inference for high-dimensional least absolute deviation (LAD) regression or median regression, as a prototype of quantile regression, under the distributed setting. Namely, we assume that data vectors $(y_j, \boldsymbol{X}_j^T)^T$ $(j = 1, \ldots, N)$ are i.i.d. and obey model,

$$y_j = \boldsymbol{X}_j^T \boldsymbol{\beta}_0 + \epsilon_j,$$

where $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ is a slope vector of interest which we will assume to be sparse. The error term $\epsilon_j$ is independent of $\boldsymbol{X}_j$ and satisfies the zero median condition $\mathbb{P}(\epsilon_j \leq 0) = 1/2$. We further assume that $N$ data are stored on $m$ machines, and each machine possesses $n = N/m$ data for simplicity. Although we focus here on a homoscedastic median regression model, our methods and theory can be generalized to arbitrary quantile levels and heteroscedastic models; cf. Remark 8.

Drawing inference on a single coordinate of $\boldsymbol{\beta}_0$ on a single machine has been explored by Belloni et al. (2015b) who propose debiased estimators assuming an auxiliary linear model for the covariates. In particular, we focus on the one-step debiased estimator considered in Belloni et al. (2015b) (see Section 3.2 for more details). A straightforward extension of such an estimator to the distributed setting is the *one-shot estimator* computed by averaging the local one-step estimators from individual machines. However, it has been pointed out in the context of linear regression that such one-shot aggregation has several drawbacks; in particular, it requires restrictive sample complexity assumptions as the simple aggregation fails to reduce the bias Jordan et al. (2019); Wang et al. (2017a). To alleviate such assumptions, we consider a multi-round aggregated one-step estimator based on new multi-round distributed quantile regression and heteroscedastic linear regression estimators (see Algorithm 6 for detailed descriptions) both of which can be solved efficiently. In particular,

the optimization problems in the multi-round quantile regression can be formulated as linear programming problems; cf. Remark 9. We show that the multi-round distributed quantile and linear regression estimators achieve near oracle convergence rates and desirable sparsity properties similar to a centralized procedure that utilizes all $N$ data after a few rounds while keeping the computational and communication costs low. Instead of solving $m$ local optimization problems of size $n$, the multi-round algorithms only solve one size-$n$ local optimization and incur a communication cost of $O(ms)$ where $s$ is the sparsity of $\boldsymbol{\beta}_0$ at each round. We also show that the proposed estimator requires much weaker sample complexity conditions than the one-shot estimator for valid inference. Specifically, under the bounded design, the multi-round estimator requires $m = O(ns^{-3}\log^{-1} n)$ while the one-shot estimator would require $m = o\big(n^{1/2}(s\log p)^{-3/2}\big)$ (more details are discussed after Corollary 3). The proposed multi-round quantile regression and linear regression estimators, which are building blocks of our multi-round one-step estimator, are inspired by Shamir et al. (2014) and Wang et al. (2017a). However, their results do not cover both estimators in the present paper and the theoretical development of our estimators requires significantly different technical analyses. We refer to the literature review in the next section for more detailed comparisons.

We further consider simultaneous confidence intervals for multiple coordinates of $\boldsymbol{\beta}_0$, for instance, $\beta_{0,1}, \cdots, \beta_{0,p_1}$, using the proposed multi-round estimator $(\hat{\beta}_k)_{k=1}^{p_1}$, where we allow $p_1$ to grow with the sample size $N$. To this end, we first show that $\hat{\beta}_k - \beta_{0k}$ can be approximated by the linear term $N^{-1}\sum_{j=1}^{N}\psi(U_j, v_{(k)j})$ uniformly for $1 \le k \le p_1$, where $\psi$ is the influence function, $v_{(k)1}, \cdots, v_{(k)N}$ are the residuals of the auxiliary regression model for the coefficient $\beta_{0,k}$, and $U_1, \ldots, U_n$ are uniform random variables on (0,1) independent

of the data. Building on high dimensional Gaussian approximation techniques developed in Chernozhukov et al. (2014, 2017a), we show that $\sqrt{N}(\hat{\beta}_k - \beta_{0k})_{k=1}^{p_1}$ can be approximated by a $p_1$-dimensional Gaussian vector uniformly over the hyperrectangles in $\mathbb{R}^{p_1}$, i.e., all sets $A$ of the form: $A = \{w \in \mathbb{R}^{p_1} : a_j \leq w_j \leq b_j \text{ for all } j = 1, \cdots, p_1\}$ for some $-\infty \leq a_j \leq b_j \leq \infty$, $j = 1, \ldots, p_1$, even when $p_1 \gg N$. As the approximating Gaussian distribution is infeasible due to its unknown covariance, we propose a novel pivotal bootstrap for simultaneous inference. Our bootstrap method is motivated by the following observation that the leading stochastic term in the prescribed expansion is conditionally pivotal in the sense that, conditionally on the data, the distribution of

$$N^{-1/2} \sum_{j=1}^{N} \psi(U_j, v_{(k)j})$$

is known up to estimable nuisance parameters. This suggests a version of bootstrap by sampling uniform random variables $U_j$ independent of the data. In practice, we replace unknown nuisance parameters by consistent estimates. We prove that the pivotal bootstrap can consistently estimate the sampling distribution of $\sqrt{N}(\hat{\beta}_k - \beta_{0k})_{k=1}^{p_1}$ uniformly over the rectangles in $\mathbb{R}^{p_1}$ even when $p_1 \gg N$. Importantly, the new pivotal bootstrap enjoys desirable scalability and computational advantages. Since the pivotal bootstrap only requires a one-time estimation of the nuisance parameters and resampling uniform random variables independently of the data, it can be easily paralleled and adapted to the distributed setting (see Algorithm 8). Further, the pivotal bootstrap is computationally more attractive than the nonparametric bootstrap as the pivotal bootstrap only requires evaluating the influence functions repeatedly while the nonparametric bootstrap requires computing the estimates many times.

From a technical perspective, the theoretical analysis of the present paper

is highly nontrivial. Our theoretical contribution mainly consists of: 1) deriving the convergence rates and sparsity properties of the proposed multi-round estimators; 2) establishing a high-dimensional Gaussian approximation to our estimate; 3) proving the validity of the new pivotal bootstrap. Each of these results relies on modern empirical process theory and high-dimensional Gaussian approximation techniques developed by Chernozhukov et al. (2014, 2017a).

In summary, this paper makes the following contributions to the literature. First, we propose a new multi-round estimator for distributed high-dimensional LAD regression that requires a weak sample complexity condition while maintaining low computation and communication costs. Second, we develop a new pivotal bootstrap for simultaneous inference of our estimator and establish its theoretical validity allowing the number of the inferred coordinates possibly exceeding the sample size. Particularly, the new pivotal bootstrap is built on an insight into the specific structure of our estimator and is well suited to modern parallel and distributed computing architectures.

### 3.1.2 Literature review

Recent years have witnessed a quickly growing literature on statistical estimation and inference in distributed environments. The mainstream research has been devoted to applying the aforementioned one-shot approach to various statistical problems including (but are not limited to) sparse linear regression Jordan et al. (2019); Wang et al. (2017a), quantile regression Volgushev et al. (2019); Chen et al. (2019, 2020), non-standard problems Banerjee et al. (2019); Shi et al. (2018), principal component analysis Fan et al. (2019). The one-shot approach

is popular since it can be easily implemented. However, it suffers from several drawbacks as pointed out by Jordan et al. (2019) that hinders further inference tasks. In particular, the one-shot approach usually requires restrictive sample complexity conditions, i.e., conditions on $m$ or $n$, to achieve valid inference. Such a condition limits its application to some important modern situations such as large sensor networks and streaming data Chen et al. (2019).

Several recent papers study alternative multi-round methods to relax the strong sample complexity condition imposed by the one-shot approach. Chen et al. (2021) propose a multi-round PCA procedure which removes the assumption on the number of machines in the previous one-shot approach. Jordan et al. (2019) consider a parametric likelihood framework and propose an iterative method to approximate the global likelihood function via multi-round aggregation. Wang et al. (2017a) also propose a similar approach for more general smooth $M$-estimation problems. These two methods are built on the distributed approximate Newton algorithm by Shamir et al. (2014) and require twice-differentiability of the loss function. Therefore, they are not directly applicable to the quantile regression considered in this paper. Our multi-round aggregation approach is also inspired by Shamir et al. (2014) but developing the statistical guarantees of resulting estimators for non-smooth problems such as quantile regression requires substantially different theoretical analyses and is much more involved. Specifically, the proof of Wang et al. (2017a) is based on the Taylor expansions of the loss function and its derivatives which are not applicable in the quantile regression case. Instead, we need to use new decompositions and deal with empirical process terms directly that require modern empirical process techniques such as local maximal inequalities and Talagrand's concentration inequality, cf. Lemmas 2 and 21 in the Appendix. Additionally,

we also prove the sparsity properties of our multi-round estimators that are not covered in the previous literature. Chen et al. (2019) propose a multi-round distributed quantile regression estimator using a smoothed quantile loss function. However, our work is substantially different from theirs. First, we do not employ any smoothing technique and therefore can avoid a delicate task of selecting a bandwidth used in smoothed quantile regression. Second, contrary to the high-dimensional setting considered in this paper, the theory of Chen et al. (2019) requires the dimension of the problem far less than the sample size. Another recent work by Battey et al. (2021) also studies the distributed quantile regression based on a new smoothing technique. Similarly, their method requires delicate choices of smoothing parameters and is focused on the low-dimensional setting for inference. Recently, Yu et al. (2021) study simultaneous inference using bootstrap for general high-dimensional $M$-estimators under the distributed setting. However, similar to Wang et al. (2017a), they also require the loss function to be sufficiently smooth. Thus, their method and results are not applicable to the quantile regression case. Further, our pivotal bootstrap is different from the multiplier bootstrap studied by Yu et al. (2021) that is one of our major contributions.

This paper also builds on the quantile regression literature. Starting from the seminal work of Koenker and Bassett (1978), quantile regression has been widely used in many real applications. We refer the readers to Koenker (2005, 2017) for a comprehensive review of quantile regression. In particular, the proposed multi-round one-step estimator is built on the work by Belloni et al. (2015b) who study pointwise inference based on the asymptotic normality for high-dimensional LAD regression in the non-distributed setting. In contrast, we develop valid distributed simultaneous inference that is applicable when

the number of inferred coordinates is increasing with or even larger than the sample size based on a new pivotal bootstrap. The proposed pivotal bootstrap is related to Parzen et al. (1994); Chernozhukov et al. (2009); He (2017); Belloni et al. (2019a); Zhang et al. (2021) who study resampling-based inference methods that build on (conditionally) pivotal influence functions in the quantile regression setup. However, their scopes and methods are substantially different from ours.

### 3.1.3 Notations

For a vector $\boldsymbol{\beta} \in \mathbb{R}^p$, let $\|\boldsymbol{\beta}\|_0$ denote the $\ell_0$-norm, i.e., the the cardinality of the nonzero coordinates of $\boldsymbol{\beta}$, and $\|\boldsymbol{\beta}\|_q := \left(\sum_{k=1}^p |\beta_k|^q\right)^{1/q}$ be the $\ell_q$-norm of $\boldsymbol{\beta}$ for $q \geq 1$. Given a set of indexes $\mathcal{I}$, we let $\boldsymbol{\beta}_{\mathcal{I}}$ denote the subvector of $\boldsymbol{\beta}$ with coordinate indexes in $\mathcal{I}$. Similarly, for a matrix $M$, $M_{\mathcal{I}}$ denotes the submatrix of $M$ with both row and column indexes in $\mathcal{I}$. We use $\mathbb{E}_n$ to abbreviate the notation $n^{-1}\sum_{j=1}^n$ and $\mathbb{E}[\cdot]$ to denote the expectation operator. We use $C$ to denote constants that are independent of both $n$ and iteration $t$ but may vary from line to line. For two sequences of positive constants $\{a_n\}_{n=1}^\infty$ and $\{b_n\}_{n=1}^\infty$, we denote $a_n = O(b_n)$ if there exists a constant $C$ independent of both $n$ and iteration $t$ such that $a_n \leq Cb_n$.

## 3.2 Inference via one-step estimator

Suppose that we observe $n$ i.i.d. data $(y_j, \boldsymbol{X}_j)^T$ ($j = 1, \ldots, N$) with $y_j \in \mathbb{R}$ and $\boldsymbol{X}_j = (X_{j1}, \cdots, X_{jp})^T \in \mathbb{R}^p$, which obey the following median regression model

$$y_j = \boldsymbol{X}_j^T \boldsymbol{\beta}_0 + \epsilon_j, \tag{3.1}$$

where $\boldsymbol{\beta}_0 = (\alpha_1, \beta_{0,2}, \cdots, \beta_{0,p})^T$ is the slope vector. Here we are interested in inference on the first coordinate of $\boldsymbol{\beta}_0$, namely $\alpha_1$. The error term $\epsilon_j$ is independent of $\boldsymbol{X}_j$ and satisfies $\mathbb{P}(\epsilon_j \leq 0) = 1/2$. We further assume that $\epsilon$ admits a density function $f_\epsilon$ such that $f_\epsilon(0) > 0$. We have in mind that the number of covariates, $p$, is large ($p \gg N$), but assume that $\boldsymbol{\beta}_0$ is sparse. Standard regularized regression estimators, such as $\ell_1$-penalized median regression, can be used to estimate $\boldsymbol{\beta}_0$, which is known to achieve near oracle convergence rates Belloni and Chernozhukov (2011). However, the bias introduced by the regularization creates a challenge for inference that is well-known in the high-dimensional statistics literature Chernozhukov et al. (2016a); Javanmard and Montanari (2018); Van de Geer et al. (2014); Zhang and Zhang (2014). To overcome this problem, Belloni et al. (2015b) propose to "debias" the regularized estimator of $\alpha_1$ using an orthogonal score equation based on the following auxiliary model,

$$X_{j1} = \boldsymbol{X}_{j(-1)}^T \boldsymbol{\theta}_0 + v_j, \qquad \mathbb{E}[v_j \mid \boldsymbol{X}_{j(-1)}] = 0 \quad (j = 1, \ldots, n), \tag{3.2}$$

where $\boldsymbol{X}_{j(-1)} = (X_{j2}, \cdots, X_{jp})^T$ and $\boldsymbol{\theta}_0$ is assumed to be sparse. Belloni et al. (2015b) show asymptotic normality of the debiased estimator for $\alpha_1$, which can be applied to inference on $\alpha_1$ that is robust to the bias induced by the regularization. Belloni et al. (2015b) also propose a one-step estimator (see Algorithm 4) but only provide brief discussions. It is worthwhile to note that in Belloni et al. (2015b), $f_\epsilon(0)$ is assumed to be known that is rarely true in practice. We

avoid this assumption in Algorithm 4 and use a kernel estimator of $f_\epsilon(0)$ instead. Belloni et al. (2015b) propose to use $\ell_1$-penalized or post $\ell_1$-penalized median regression in the first step and Lasso or post-Lasso regression in the second step.

Such a one-step estimator is attractive in the distributed setting due to its arithmetic average nature in its explicit form that allows for a direct parallel computation (see Algorithm 6). Thus, we shall adapt the one-step estimator to the distributed setting. Specifically, we propose a new algorithm (Algorithm 6) tailored for computing the one-step estimator in the distributed setting that can achieve a good balance between computational and communication costs where we use multi-round quantile regression and heteroscedastic Lasso studied in Section 3.3.1 to relax the constraint on the number of machines required by a naive one-shot distributed estimator. For inference, we propose a new pivotal bootstrap for valid simultaneous inference and provide a computationally efficient algorithm (Algorithm 8) for our pivotal bootstrap that takes advantage of the parallel computing in the distributed setting.

**Remark 8** (Generalizations to arbitrary quantiles and heteroscedastic models)**.** We consider the median regression for the demonstration purpose and in fact, our methods are valid for arbitrary quantile levels as long as we assume a homoscedastic model, i.e., $\epsilon$ is independent of $\boldsymbol{X}$. Belloni et al. (2019b) generalize the above one-step estimator to the heteroscedastic median regression model. To accommodate for the heteroscedasticity in our methods, we can adapt the heteroscedastic version of the one-step estimator proposed by Belloni et al. (2019b) to our framework. However, the heteroscedastic one-step estimator requires accurate estimates of the conditional density $f_{\epsilon_j}(0 \mid \boldsymbol{X}_j)$ for all $1 \leq j \leq n$, which is practically hard in the current distributed setting and therefore con-

**Algorithm 4:** One-step estimator in Belloni et al. (2015b)

**Input**: Data: $\{(y_j, \boldsymbol{X}_j)\}$ $(1 \leq j \leq n)$; Median score function: $\varphi(t) = 1/2 - I(t \leq 0)$; Kernel function: $K(t)$; Smoothing bandwidth: $h$.

**Step 1**. Run high-dimensional median regression of $y_j$ on $\boldsymbol{X}_j$ to compute the estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}_0$; keep the residual $\hat{\epsilon}_j = y_j - \boldsymbol{X}_j^T \hat{\boldsymbol{\beta}}_0$ and $\hat{\alpha}_1$, the initial biased estimator of $\alpha_1$ obtained from $\hat{\boldsymbol{\beta}}$.

**Step 2**. Run high-dimensional linear regression of $X_{j1}$ on $\boldsymbol{X}_{j(-1)}$ to compute the estimator $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}_0$; keep the residual $\hat{v}_j = X_{j1} - \boldsymbol{X}_{j(-1)}^T \hat{\theta}$.

**Step 3**. Estimate $f_\epsilon(0)$ by $\hat{f}_\epsilon(0) = n^{-1} \sum_{j=1}^n h^{-1} K(\hat{\epsilon}_j / h)$.

**Step 4**. Form the one-step estimator as follows,

$$\check{\alpha}_1 = \hat{\alpha}_1 + \big[\mathbb{E}_n[\hat{f}_\epsilon(0)\hat{v}_j^2]\big]^{-1} \mathbb{E}_n\big[\varphi(\hat{\epsilon}_j)\hat{v}_j\big]. \tag{3.3}$$

strains its applications in practice. Besides, the analysis of the impact of such estimation is also delicate which is out of the scope of this paper and we leave it for future research.

## 3.3 Main results

In the following, we will assume a distributed high-dimensional setting: the data are stored in $m$ machines and for $1 \leq i \leq m$, the $i$-th machine has access to i.i.d. observations $\{(\boldsymbol{X}_{ij}, y_{ij})\}_{j=1}^n$ with $\boldsymbol{X}_{ij} \in \mathbb{R}^p$, $y_{ij} \in \mathbb{R}$ and $p \gg n$. We further assume that the data stored on different machines are independent with each other and the number of the machines is increasing with the sample size, i.e., $m \to \infty$ when $n \to \infty$. Our main results consist of two parts. First, we will present a multi-round estimation framework for a general penalized $M$-estimation problem and consider two important examples: quantile regression and linear regression, both of which will be used for computing the distributed

one-step estimator. In the second part, we focus on inference for the distributed one-step estimator. We will first investigate the pointwise inference and demonstrate the advantage of using our multi-round one-step over the straightforward one-shot estimator. Then, we present our pivotal bootstrap for the simultaneous inference of the multi-round distributed one-step estimator together with statistical guarantees.

### 3.3.1 Distributed estimation via multi-round aggregation

In this section, we present multi-round aggregation algorithms for quantile regression and heteroscedastic linear regression together with their statistical guarantees. We will first informally introduce a distributed multi-round estimation framework for general penalized $M$-estimation problems and then treat the two particular regression problems above as specific examples.

Suppose that we are interested in a parameter $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ attached to the marginal distribution of the observations $\{(\boldsymbol{X}_{ij}, y_{ij})\}_{j=1}^n$ $(1 \leq i \leq m)$. Let $\ell(\cdot)$ be a convex loss function such that $\boldsymbol{\beta}_0$ is a minimizer of the population risk, i.e.,

$$\boldsymbol{\beta}_0 \in \arg\min_{\boldsymbol{\beta}} \mathbb{E}_{(\boldsymbol{X},y)}[\ell(\boldsymbol{\beta}, \boldsymbol{X}, y)].$$

Since we are interested in a high-dimensional setting, we assume that $\boldsymbol{\beta}_0$ is $s$-sparse i.e., $\|\boldsymbol{\beta}_0\|_0 = s$ and define the centralized penalized loss function as

$$\mathcal{L}(\boldsymbol{\beta}) := \sum_{i=1}^m \mathcal{L}_i(\boldsymbol{\beta}) + \mathcal{P}(\boldsymbol{\beta}), \qquad \mathcal{L}_i(\boldsymbol{\beta}) := \frac{1}{n}\sum_{j=1}^n \ell(\boldsymbol{\beta}, \boldsymbol{X}_{ij}, y_{ij}), \tag{3.4}$$

where $\mathcal{P}(\cdot)$ is a penalty function used to deal with the high dimensionality. Ideally, we can minimize the centralized loss function to estimate $\boldsymbol{\beta}_0$. However, in

practice, the memory constraints and the distributed nature of the data make the above centralized loss function infeasible.

To achieve estimation performance that is comparable to the oracle estimator based on minimizing the centralized loss function, we propose the following multi-round estimation procedure inspired by Wang et al. (2017a). Suppose that, at the $(t+1)$-th iteration, the master machine (for instance, machine 1) computes the estimator by solving the following optimization problem,

$$\hat{\boldsymbol{\beta}}_{t+1} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \left\{ \mathcal{L}_1(\boldsymbol{\beta}) + \left\langle \frac{1}{m}\sum_{i=1}^{m} \nabla\mathcal{L}_i(\hat{\boldsymbol{\beta}}_t) - \nabla\mathcal{L}_1(\hat{\boldsymbol{\beta}}_t), \boldsymbol{\beta} \right\rangle + \mathcal{P}_{t+1}(\boldsymbol{\beta}) \right\}, \quad (3.5)$$

where $\langle\cdot,\cdot\rangle$ denotes the Euclidean inner product, $\hat{\boldsymbol{\beta}}_t$ is the estimator at the $t$-th iteration and $\nabla\mathcal{L}_i(\cdot)$ is the gradient or subgradient of $\mathcal{L}_i(\cdot)$. See Algorithm 5 for the detailed description of the procedure. In Algorithm 5, the initial estimator

**Algorithm 5:** General multi-round aggregation algorithm

**Input:** Data: $\{(\boldsymbol{X}_{ij}, y_{ij})\}$ $1 \le i \le m$, $1 \le j \le n$; initial estimator $\hat{\boldsymbol{\beta}}_{\text{ini}}$.
**for** $t = 0, 1, 2, \cdots$ **do**
  Machine 1 passes $\hat{\boldsymbol{\beta}}_t$ ($\hat{\boldsymbol{\beta}}_{\text{ini}}$ if $t = 0$) to the rest of machines.
  **for** $2 \le i \le m$ **do**
    Machine $i$ calculates $\nabla\mathcal{L}_i(\hat{\boldsymbol{\beta}}_t)$ and passes it to Machine 1.
  **end**
  Machine 1 computes $\hat{\boldsymbol{\beta}}_{t+1}$ by solving (3.5).
**end**

$\hat{\boldsymbol{\beta}}_{\text{ini}}$ can be taken as the estimator computed by minimizing the local penalized loss function on machine 1 if no existing estimator is available. In the following two sections, we will apply the above general multi-round algorithm to two important cases: quantile regression and heteroscedastic linear regression, both of which will be building blocks for our multi-round one-step estimator.

Wang et al. (2017a) study a similar multi-round procedure but assume that the loss function $\ell(\cdot)$ is at least three times differentiable. Thus, their results do

not cover the quantile regression case. As discussed in the literature review, the theoretical development for the multi-round quantile regression requires substantially different technical tools and analyses that are much more involved than the smooth loss function case. Also, the convergence rate results for the multi-round sparse heteroscedastic linear regression and the sparsity properties of both the multi-round quantile regression and linear regression estimators are new. In particular, the sparsity properties of these two multi-round estimators are essential for the inference of the multi-round one-step estimator in Section 3.3.2.

**Adaptively weighted quantile regression**

In this section, we investigate the distributed multi-round estimation for quantile regression. Specifically, we focus on the following linear quantile regression model,

$$y = \boldsymbol{X}^T\boldsymbol{\beta}_0 + \epsilon \quad \text{and} \quad \mathbb{P}(\epsilon < 0 \mid \boldsymbol{X}) = \tau,$$

where $\tau$ is the quantile index of interest. For the quantile regression model, we will use the quantile loss function, i.e. $\ell(\boldsymbol{X}, y, \boldsymbol{\beta}) = \{\tau - I(y \leq \boldsymbol{X}^T\boldsymbol{\beta})\}(y - \boldsymbol{X}^T\boldsymbol{\beta})$. Hence the local loss function iis $\mathcal{L}_i(\boldsymbol{\beta}) := n^{-1}\sum_{j=1}^{n}\{\tau - I(y_{ij} \leq \boldsymbol{X}_{ij}^T\boldsymbol{\beta})\}(y_{ij} - \boldsymbol{X}_{ij}^T\boldsymbol{\beta})$ and the local loss subgradient is $\nabla\mathcal{L}_i(\boldsymbol{\beta}) = n^{-1}\sum_{j=1}^{n}\{I(y_{ij} \leq \boldsymbol{X}_{ij}^T\boldsymbol{\beta}) - \tau\}\boldsymbol{X}_{ij}$ for $1 \leq i \leq m$. Formally, we consider the following adaptively $\ell_1$-penalized optimization problem at the $(t+1)$ iteration,

$$\hat{\boldsymbol{\beta}}_{t+1} := \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min}\left\{\mathcal{L}_1(\boldsymbol{\beta}) + \left\langle \frac{1}{m}\sum_{i=1}^{m}\nabla\mathcal{L}_i(\hat{\boldsymbol{\beta}}_t) - \nabla\mathcal{L}_1(\hat{\boldsymbol{\beta}}_t), \boldsymbol{\beta}\right\rangle + \lambda_{t+1}\sum_{k=1}^{p}\omega_{t+1,k}|\beta_k|\right\}, \tag{3.6}$$

where $\lambda_{t+1}$ is the penalty level and $\omega_{t+1,k}$ is the adaptive weight for $\beta_k$, both of which are allowed to change with iterations. Several adaptive weights $\omega$ have

been proposed in the literature Fan et al. (2014); Zheng et al. (2015) among which we will use $\omega_{t+1,k} = 1/|\hat{\beta}_{k,t}|$ in our numerical analysis.

**Remark 9** (Computational aspect)**.** From equation (3.6), we see that the multi-round quantile regression optimization problem at each iteration can be formulated as a linear programming (LP) problem and hence can be solved efficiently, cf. Appendix C.8. There are many efficient LP solvers available. In our simulation study, for each iteration, we first formulate (3.6) as an LP problem and solve it using *Mosek* ApS (2021), which is a state of the art LP solver.

Now, we present the theoretical guarantees for our multi-round quantile regression estimator. If we denote the support of $\boldsymbol{\beta}_0$ by $\mathcal{S}$ and define the oracle parameter space $\mathbb{R}_{\mathcal{S}} := \{\delta \in \mathbb{R}^p : \delta_{\mathcal{S}^c} = 0\}$, we define the following oracle estimator at the $(t+1)$-th iteration,

$$\hat{\boldsymbol{\beta}}^{\circ}_{t+1} := \underset{\boldsymbol{\beta} \in \mathbb{R}_{\mathcal{S}}}{\arg\min} \left\{ \mathcal{L}_1(\boldsymbol{\beta}) + \left\langle \frac{1}{m} \sum_{i=1}^{m} \nabla \mathcal{L}_i(\hat{\boldsymbol{\beta}}_t) - \nabla \mathcal{L}_1(\hat{\boldsymbol{\beta}}_t), \boldsymbol{\beta} \right\rangle + \lambda_{t+1} \sum_{k \in \mathcal{S}} \omega_{t+1,k} |\beta_k| \right\}. \tag{3.7}$$

We will show that under mild conditions, $\hat{\boldsymbol{\beta}}_{t+1} = \hat{\boldsymbol{\beta}}^{\circ}_{t+1}$ with high probability, which not only shows the model selection consistency of the multi-round estimator, but also implies that the multi-round estimator can achieve the same convergence rate as the oracle estimator.

We first present the assumptions. We denote the support of the covariate vector $\boldsymbol{X}$ by $\mathcal{X}$. Define the Gram matrix $\boldsymbol{\Sigma} = \mathbb{E}[\boldsymbol{X}\boldsymbol{X}^T]$ and sparse sphere $\mathcal{R}_{\mathcal{S}} := \{\boldsymbol{\delta} \in \mathbb{R}^p : \|\boldsymbol{\delta}\| = \|\boldsymbol{\delta}_{\mathcal{S}}\| = 1\}$. Let $f(t \mid \boldsymbol{x})$ denotes the conditional density function of $Y - \boldsymbol{X}^T\boldsymbol{\beta}_0$ given $\boldsymbol{X}$. To facilitate the theoretical analysis for the multi-round adaptively weighted quantile regression, we make the following assumption.

**Assumption 4** (Model assumption)**.** *(i)* $\|\boldsymbol{X}\|_\infty \leq K_n$*; (ii) The conditional density* $f(\cdot \mid \boldsymbol{x})$ *is continuously differentiable and* $f(t \mid \boldsymbol{x}) \leq \bar{f}$*,* $f(0 \mid \boldsymbol{x}) \geq \underline{f}$ *and* $f'(t \mid$

$\boldsymbol{x}) \leq \bar{f}'$ *for all* $\boldsymbol{x} \in \mathcal{X}$ *and* $t \in \mathbb{R}$*, where* $\bar{f}$*,* $\underline{f}$ *and* $\bar{f}'$ *are finite positive constants independent of* $n$*; (iii)* $\boldsymbol{\Sigma}_{\mathcal{S}}$ *has bounded eigenvalues from above and below, i.e.,* $0 < \lambda_{\min} \leq \inf_{\boldsymbol{\delta} \in \mathcal{R}_{\mathcal{S}}} \boldsymbol{\delta}_{\mathcal{S}}^T \boldsymbol{\Sigma}_{\mathcal{S}} \boldsymbol{\delta}_{\mathcal{S}} \leq \sup_{\boldsymbol{\delta} \in \mathcal{R}_{\mathcal{S}}} \boldsymbol{\delta}_{\mathcal{S}}^T \boldsymbol{\Sigma}_{\mathcal{S}} \boldsymbol{\delta}_{\mathcal{S}} \leq \lambda_{\max} < \infty$ *where* $\lambda_{\min}$ *and* $\lambda_{\max}$ *are constants independent of* $n$*; (iv)* $\sup_{k \in \mathcal{S}^c, \boldsymbol{\delta} \in \mathcal{R}_{\mathcal{S}}} \mathbb{E}[(X_k \boldsymbol{X}^T \boldsymbol{\delta})^2] = O(a_n^2)$.

Conditions (i)–(iv) are (more or less) standard in the (high-dimensional) quantile regression literature Koenker (2005). Condition (i) assumes the covariates to be bounded while the upper bound is allowed to increase with $n$. Such a condition is common in high-dimensional quantile regression analysis Fan et al. (2014); Zheng et al. (2015) and will be used to apply Talagrand's concentration inequality (Lemma 21). It can be generalized to the unbounded subgaussian design by using a corresponding version of Talagrand's inequality, cf. Theorem 4 in Adamczak (2008). Condition (ii) imposes mild conditions on the conditional density of the response given the covariates that are prevalent in the theoretical analysis of quantile regression Belloni and Chernozhukov (2011); Zheng et al. (2015). Condition (iii) concerns the eigenvalues of the Gram matrix of the relevant variables and is common in the high-dimensional statistics literature. Condition (iv) essentially characterizes the correlation between the relevant and irrelevant variables and will be needed to establish the model selection consistency of the estimator. Similar conditions can be seen in the variable selection literature for quantile regression, for instance, Theorem 3.3 in Zheng et al. (2015). In particular, if $\boldsymbol{X} \sim N(0, \boldsymbol{\Sigma})$ for some $\boldsymbol{\Sigma}$ satisfying $\max_{k \in \mathcal{S}^c} \boldsymbol{\Sigma}_{kk} = O(1)$ and Condition (iii) above is satisfied, then a simple application of Cauchy-Schwarz inequality shows that $\sup_{k \in \mathcal{S}^c, \boldsymbol{\delta} \in \mathcal{R}_{\mathcal{S}}} \mathbb{E}[(X_k \boldsymbol{X}^T \boldsymbol{\delta})^2] = O(1)$.

Additionally, we define

$$q := \inf_{\boldsymbol{\delta} \in \mathcal{R}_{\mathcal{S}}} \frac{\mathbb{E}[(\boldsymbol{X}^T \boldsymbol{\delta})^2]^{3/2}}{\mathbb{E}\left[\left|\boldsymbol{X}^T \boldsymbol{\delta}\right|^3\right]},$$

which is the restricted nonlinear impact (RNI) coefficient that controls the quality of minoration of the quantile loss function by a quadratic function over the true model Belloni and Chernozhukov (2011). Under Conditions (i)–(iii), it is not difficult to see that $q \geq C(K_n\sqrt{s})^{-1} > 0$ for some constant $C > 0$ independent of $n$.

In addition to the model level assumption above, we also need the following assumption for the algorithm implementation. We denote by $\hat{\boldsymbol{\beta}}_{\text{ini}}$ the initial estimator of $\boldsymbol{\beta}_0$. Also let

$$\eta_t = \frac{s^2 K_n^2 \log n}{n} \cdot \left(\sqrt{\frac{n}{K_n^2 \log n}}\right)^{\frac{1}{2^t}} \bigvee \sqrt{\frac{sK_n^2 \log n}{N}}.$$

**Assumption 5** (Algorithm implementation)**.**

*(i)* $\|\hat{\boldsymbol{\beta}}_{\text{ini}} - \boldsymbol{\beta}_0\| = O\big(\sqrt{sK_n^2 \log p/n}\big)$ *and* $\|\hat{\boldsymbol{\beta}}_{\text{ini}}\|_0 = O(s)$ *with probability at least* $1 - \gamma_n$.

*(ii) The penalty parameters at the* $(t+1)$*-th iteration satisfy*

- $\max_{k\in\mathcal{S}} \lambda_{t+1}\omega_{t+1,k} = O\big(\sqrt{\eta_t s K_n^2 \log n/n} \vee \sqrt{K_n^2 \log n/N}\big)$;
- $(\min_{k\in\mathcal{S}^c} \omega_{t+1,k})^{-1}\eta_t^{-1} = O(1)$ *and* $K_n\sqrt{s\log n}\eta_t\big((a_n\sqrt{s} \vee 1)\sqrt{\eta_t/n} \vee N^{-1/2}\big)\lambda_{t+1}^{-1} = o(1)$.

Condition (i) concerns with the convergence rate and sparsity property of the initial estimator $\hat{\boldsymbol{\beta}}_{\text{ini}}$ and can be achieved by using the $\ell_1$-penalized quantile regression Belloni and Chernozhukov (2011) or the adaptively $\ell_1$-penalized quantile regression estimator studied by Zheng et al. (2015) on a dataset of size $O(n)$. Condition (ii) concerns the penalty level of each iteration. Particularly, if we take $\omega_{t+1,k} = |\beta_k|^{-1}$, then the first condition in Condition (ii) can be interpreted as a signal strength condition. Similar conditions appear in Zheng et al.

(2015), Fan et al. (2014) and other adaptively penalized quantile regression literature. In practice, we recommend a "cross-validation" type procedure to select the penalty parameters in a data adaptive manner; cf. Section 3.4.1.

Under the above assumptions, the following theorem characterizes the convergence rate and the model selection consistency of the multi-round quantile regression estimator. Recall $C$ denotes a constant independent of both $n$ and iteration $t$ but may vary from place to place.

**Theorem 6.** *Suppose Assumptions 4 and 5 hold and assume further that*

$$m = O\Big(\frac{n}{s^3 K_n^2 \log n}\Big), \quad p = o(n^s) \quad \textit{and} \quad C\sqrt{\frac{s^2 K_n^2 \log n}{n}} \cdot \eta_0^{1/2} \leq q. \tag{3.8}$$

*Then after at least* $t_{\text{opt}} = \lceil \log_2 \big( \log a / \log b \big) \rceil$ *iterations, i.e.,* $t \geq t_{opt}$*, with* $a = \frac{n}{K_n^2 \log n}$ *and* $b = \frac{n}{ms^3 K_n^2 \log n}$*, we have*

$$\|\hat{\boldsymbol{\beta}}_t - \boldsymbol{\beta}_0\| \leq C\sqrt{sK_n^2 \log n / N} \quad \textit{and} \quad \hat{\boldsymbol{\beta}}_t = \hat{\boldsymbol{\beta}}_t^{\circ}$$

*with probability at least* $1 - Ctpn^{-s} - \gamma_n$.

Condition (3.8) allows for ultra-high dimension of $p$, as $s$ is allowed to increase with $n$, $s = s_n \to \infty$ as $n \to \infty$. More importantly, it will be shown in Section 3.3.2 that the condition on the number of machines ($m$) is less stringent when the multi-round algorithm is used for inference compared with the one-shot method. The proof of Theorem 6 consists of two steps: 1) we prove a delicate recursive bound the convergence rate for the oracle estimator $\hat{\boldsymbol{\beta}}^{\circ}$ building on modern empirical process theory; 2) then we show that $\hat{\boldsymbol{\beta}}_t = \hat{\boldsymbol{\beta}}_t^{\circ}$ with high probability and prove the theorem via a delicate recursive argument. In the numerical analysis, we set $\omega_{t+1,k} = |\hat{\beta}_{t,k}|^{-1}$ though we assume deterministic $\omega_{t+1,k}$ in Theorem 6. The corollary below shows that the results of Theorem 6

continue to hold under such a choice of $\omega$. Our numerical results also show a promising performance by using such a choice.

**Corollary 2.** *Assume (i)* $\max_{k\in\mathcal{S}} \lambda_{t+1}|\beta_{0k}|^{-1} = O\big(\sqrt{\eta_t s K_n^2 \log n/n} \vee \sqrt{K_n^2 \log n/N}\big)$, $\max_{k\in\mathcal{S}} |\beta_{0k}|^{-1}\big(\sqrt{sK_n^2 \log p/n}\big) = o(1)$; *(ii)* $(\min_{k\in\mathcal{S}^c} |\hat{\beta}_{\mathrm{ini},k}|^{-1})^{-1}\eta_0^{-1} = O(1)$; *(iii) other conditions that do not involve* $\omega$ *in Theorem 6. If we take* $\omega_{t+1,k} = |\hat{\beta}_{t,k}|^{-1}$, *the conclusions of Theorem 6 continue to hold.*

**Heteroscedastic Lasso**

In this section, we investigate the distributed multi-round estimation for the high-dimensional heteroscedastic linear regression. Specifically, we focus on the following model

$$y = \boldsymbol{X}^T\boldsymbol{\beta}_0 + \epsilon \quad \text{and} \quad \mathbb{E}[\epsilon \mid \boldsymbol{X}] = 0.$$

The previous literature on the multi-round sparse linear regression requires $\epsilon$ to be independent of $\boldsymbol{X}$ Wang et al. (2017a) which is inappropriate for the application to our one-step estimator where we assume a heteroscedastic linear model, cf. equation (3.2). Besides, we establish the sparsity properties of the multi-round heteroscedastic lasso estimator that is also new in the literature.

For the linear model, we use the square loss function $\ell(\boldsymbol{X}, y, \boldsymbol{\beta}) = \big(y - \boldsymbol{X}^T\boldsymbol{\beta}\big)^2$. Hence we have the local loss function $\mathcal{L}_i(\boldsymbol{\beta}) := n^{-1}\sum_{j=1}^n \big(y_{ij} - \boldsymbol{X}_{ij}^T\boldsymbol{\beta}\big)^2$ and the local loss gradient $\nabla\mathcal{L}_i(\boldsymbol{\beta}) = 2n^{-1}\sum_{j=1}^n \big(\boldsymbol{X}_{ij}^T\boldsymbol{\beta} - y_{ij}\big)\boldsymbol{X}_{ij}$ for $1 \le i \le m$. We consider the following $\ell_1$-penalized optimization problem at the $(t+1)$ iteration,

$$\hat{\boldsymbol{\beta}}_{t+1} := \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \left\{ \mathcal{L}_1(\boldsymbol{\beta}) + \left\langle \frac{1}{m}\sum_{i=1}^m \nabla\mathcal{L}_i(\hat{\boldsymbol{\beta}}_t) - \nabla\mathcal{L}_1(\hat{\boldsymbol{\beta}}_t), \boldsymbol{\beta} \right\rangle + \lambda_{t+1}\|\boldsymbol{\beta}\|_1 \right\}, \quad (3.9)$$

where $\lambda_{t+1}$ is allowed to change with iterations.

Now, we establish the convergence rate and the sparsity property of the multi-round heteroscedastic lasso estimator. Before proceeding, we introduce some notations. We denote the population Gram matrix by $\boldsymbol{\Sigma} := \mathbb{E}[\boldsymbol{X}\boldsymbol{X}^T]$ and define the minimal and maximal *$m$-sparse eigenvalues* of $\boldsymbol{\Sigma}$ as

$$\bar{\phi}_{\min}(m, \boldsymbol{\Sigma}) := \min_{1 \le \|\boldsymbol{\delta}\|_0 \le m} \frac{\boldsymbol{\delta}^T \boldsymbol{\Sigma} \boldsymbol{\delta}}{\|\boldsymbol{\delta}\|^2}, \qquad \bar{\phi}_{\max}(m, \boldsymbol{\Sigma}) := \max_{1 \le \|\boldsymbol{\delta}\|_0 \le m} \frac{\boldsymbol{\delta}^T \boldsymbol{\Sigma} \boldsymbol{\delta}}{\|\boldsymbol{\delta}\|^2},$$

where $m = 1, \ldots, p$. Let $\ell_n \to \infty$ be a sequence of positive constants. We will make the following assumption.

**Assumption 6.** *(i)* $|y| \vee \|\boldsymbol{X}\|_\infty \le K_n$ *and* $|\epsilon| \le V_n$*; (ii)* $0 < c_1 \le \bar{\phi}_{\min}(\ell_n s, \boldsymbol{\Sigma}) \le \bar{\phi}_{\max}(\ell_n s, \boldsymbol{\Sigma}) \le C_1$*; (iii)* $K_n^2 \ell_n s \log p \log^3(K_n^2 \ell_n s \log p) \vee s^2 K_n^4 \log p = o(n)$.

Condition (i) assumes that $y$, $\boldsymbol{X}$ and $\epsilon$ are bounded. In particular, when $|y| \vee \|\boldsymbol{X}\|_\infty \le K_n$, we have $|\epsilon| \le (\|\theta_0\|_1 + 1)K_n$ and therefore we may take $V_n \ge (\|\theta_0\|_1 + 1)K_n$. Bounded covariates and responses are assumed in order to be consistent with the design assumption for the multi-round quantile regression both of which will be building blocks for our multi-round one-step estimator defined in the following sections. The boundedness of $\epsilon$ is used to deal with the heteroscedasticity and the non-Gaussianity in the proof. Conditions (ii) and (iii) together guarantee that the minimal and maximal sparse eigenvalues of the *empirical Gram matrix*, $\boldsymbol{\Sigma}_n := n^{-1} \sum_{j=1}^n \boldsymbol{X}_j \boldsymbol{X}_j^T$, are bounded and bounded away from zero with high probability, i.e.,

$$0 < c_1' \le \bar{\phi}_{\min}(\ell_n s, \boldsymbol{\Sigma}_n) \le \bar{\phi}_{\max}(\ell_n s, \boldsymbol{\Sigma}_n) \le C_1',$$

with probability $1 - \nu_n$ for a sequence of positive constants $\nu_n \to 0$, cf. Rudelson and Zhou (2012). Such sparse eigenvalue conditions are common in the high-

dimensional statistics literature Bickel et al. (2009); Belloni and Chernozhukov (2011); Belloni et al. (2012); Bühlmann and Van De Geer (2011).

We also need the following implementation assumption that specifies the theoretical choice of the penalty level at each iteration.

**Assumption 7** (Algorithm implementation)**.**

*(i). For some* $\eta_0 > 0$, $\|\hat{\boldsymbol{\beta}}_{\text{ini}} - \boldsymbol{\beta}_0\| = O(\eta_0)$ *with probability at least* $1 - \gamma_n$.

*(ii). The penalty parameter at the* $(t+1)$*-th iteration* $(t \geq 0)$, $\lambda_{t+1}$, *satisfies* $C_1\Gamma_t \geq \lambda_{t+1} \geq C_2\Gamma_t$ *for some sufficiently large constants* $C_1 \geq C_2$ *independent of* $n$ *and* $t$, *and we define*

$$\Gamma_t := V_n K_n \sqrt{\frac{\log p}{N}} + K_n^2 \sqrt{\frac{\log p}{n}} \cdot \eta_t \quad \textit{and} \quad \eta_{t+1} := C(V_n K_n s \sqrt{\frac{\log p}{N}} + K_n^2 s \sqrt{\frac{\log p}{n}} \cdot \eta_t).$$

The following theorem characterizes the convergence rate and sparsity property of the multi-round heteroscedastic lasso estimator.

**Theorem 7.** *Suppose that Assumptions 6 and 7 hold. If we use the initial estimator* $\hat{\boldsymbol{\beta}}_{\text{ini}}$, *then after at least* $t_{\text{opt}} = \lceil \big(\log(b/\eta_0)\big)/\log a \rceil$ *iterations, i.e.,* $t \geq t_{opt}$, *with* $a = sK_n^2\sqrt{\log p/n}$ *and* $b = sV_nK_n\sqrt{\log p/N}$, *we have*

$$\|\hat{\boldsymbol{\beta}}_t - \boldsymbol{\beta}_0\|_1 \leq CV_nK_ns\sqrt{\log p/N} \quad \textit{and} \quad \|\hat{\boldsymbol{\beta}}_t - \boldsymbol{\beta}_0\|_2 \leq CV_nK_n\sqrt{s\log p/N},$$

*with probability* $1 - p^{-1} - \gamma_n - \nu_n$. *Additionally,* $\|\hat{\boldsymbol{\beta}}_t\|_0 = O(s)$ *with same probability.*

The common initial estimator $\hat{\boldsymbol{\beta}}_{\text{ini}}$ would be the regular lasso estimator computed on Machine 1 that achieves $\eta_0 = s\sqrt{\log p/n}$ and in this case, we can show $t_{opt} \leq C \log m$. We note that when $K_n$ and $V_n$ are independent of $n$, the multi-round estimator achieves the conventional near oracle convergence rates

$s\sqrt{\log p/N}$ and $\sqrt{s\log p/N}$ in the $\ell_1$ and $\ell_2$-norms, respectively. Similar to the quantile regression case, the proof of the convergence rate in Theorem 7 is based on a recursive argument. Once the convergence rate is proved, we can establish the sparsity property of the estimator by a similar analysis as in Belloni et al. (2012).

### 3.3.2 Distributed inference for LAD regression

In this section, we consider distributed inference for the high-dimensional LAD regression based on the one-step estimator. Although we focus on the LAD regression for presentation, the method can be generalized to quantile regression with arbitrary quantile index (see Remark 8). In the following, we first investigate the pointwise inference based on the multi-round one-step estimator. We show that our multi-round estimator imposes a less stringent condition over the number of machines compared with the one-shot approach (see Section 3.3.2 for the definition). Next, we propose a new pivotal bootstrap method for our multi-round estimator that allows for simultaneous inference on multiple coefficients whose number can be larger than the sample size $n$.

#### Pointwise inference

Recall that we assume the distributed setting that we have $m$ machines and each machine stores $n$ data, $(\boldsymbol{X}_{ij}, y_{ij})$ $(1 \le i \le m,\, 1 \le j \le n)$, that are i.i.d. generated from the following model,

$$y = \boldsymbol{X}^T\boldsymbol{\beta}_0 + \epsilon,\ \epsilon \perp\!\!\!\perp \boldsymbol{X} \quad \text{and} \quad \mathbb{P}(\epsilon \le 0) = 1/2, \tag{3.10}$$

where $\boldsymbol{\beta}_0 = (\alpha_1, \beta_{02}, \cdots, \beta_{0p})^T$, $\|\boldsymbol{\beta}_0\|_0 \leq s$ and $\alpha_1$ is the coefficient of interest. Additionally, we assume the following auxiliary model

$$X_1 = \boldsymbol{X}_{(-1)}^T \boldsymbol{\theta}_0 + v, \quad \mathbb{E}[v \mid \boldsymbol{X}_{(-1)}] = 0, \tag{3.11}$$

where we denote $\boldsymbol{X}_{(-1)} = (X_2, \cdots, X_p)^T$ and assume $\|\boldsymbol{\theta}_0\|_0 \leq s$.

Building on the analysis of the last section, we propose Algorithm 6 to compute our multi-round distributed one-step estimator. An alternative method under the distributed setting is to calculate the one-step estimator independently on each machine and then average the $m$ estimators as the final output. We will refer to such a method as *one-shot aggregation* that is described in Algorithm 7. We will consider inference for $\alpha_1$ based on the asymptotic normality of the resulting estimators from the above two approaches. It turns out that the multi-round estimator imposes less stringent conditions on the number of machines ($m$) to be used than the one-shot approach.

---

**Algorithm 6:** Multi-round one-step estimator

---

**Input**: Data: $\{(\boldsymbol{X}_{ij}, y_{ij})\}$ ($1 \leq i \leq m$, $1 \leq j \leq n$); Median score function: $\varphi(t) = 1/2 - I(t \leq 0)$; Kernel function: $K(t)$; Bandwidth: $h$.

**Step 1**. Run multi-round adaptively weighted median regression and heteroscedastic lasso among $m$ machines; Machine 1 obtains the final estimators $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}_0$ and $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}_0$; Machine 1 broadcasts sparse estimators $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\theta}}$ to the rest of machines and keeps the initial biased estimator $\hat{\alpha}_1$ obtained from $\hat{\boldsymbol{\beta}}$.

**Step 2**. For $1 \leq i \leq m$, machine $i$ computes $\mathbb{E}_n[h^{-1}K(\hat{\epsilon}_{ij}/h)]$, $\mathbb{E}_n[\varphi(\hat{\epsilon}_{ij})\hat{v}_{ij}]$ and $\mathbb{E}_n[\hat{v}_{ij}^2]$ where $\hat{\epsilon}_{ij} = y_{ij} - \boldsymbol{X}_{ij}^T\hat{\boldsymbol{\beta}}$ and $\hat{v}_{ij} = X_{ij1} - \boldsymbol{X}_{ij(-1)}^T\hat{\boldsymbol{\theta}}$, and then passes these quantities back to Machine 1.

**Step 3**. Machine 1 forms the one-step estimator as follows,

$$\check{\alpha}_1^{\mathrm{ma}} = \hat{\alpha}_1 + \big[m^{-2}\textstyle\sum_{i=1}^m \mathbb{E}_n[h^{-1}K(\hat{\epsilon}_{ij}/h)] \cdot \sum_{i=1}^m \mathbb{E}_n[\hat{v}_{ij}^2]\big]^{-1} \cdot \big[m^{-1}\textstyle\sum_{i=1}^m \mathbb{E}_n[\varphi(\hat{\epsilon}_{ij})\hat{v}_{ij}]\big].$$

---

**Algorithm 7:** One-shot aggregation

**Input:** Data: $\{(\boldsymbol{X}_{ij}, y_{ij})\}$ $1 \le i \le m$, $1 \le j \le n$.

**Step 1**. For $1 \le i \le m$, machine $i$ calculates one-step estimator $\check{\alpha}_1^{(i)}$ according to Algorithm 4; Then machine $i$ passes $\check{\alpha}_1^{(i)}$ to machine 1.

**Step 2**. Machine 1 outputs the final estimator by taking the average

$$\check{\alpha}_1^{\mathrm{oa}} = \frac{1}{m}\sum_{i=1}^{m} \check{\alpha}_1^{(i)}.$$

**Remark 10** (Computational and communication cost comparison)**.** In this remark, we compare the computation and communication costs of the multi-round and one-shot estimators. Suppose that we use the same algorithm to solve LP problems appearing in quantile regression in both estimators and denote by $T(n,p)$ the computational cost to solve an LP problem of size $n$ by $p$. We apply the proximal gradient descent Nesterov (2018) to solve high-dimensional linear regression for both estimators. Then, the computational cost of the one-shot estimator is $m(T(n,p)+O(rnp)) = O\big(mT(n,p)+mnp\big)$ where $r$ is the maximum number of iterations for the proximal gradient descent algorithm and $O(np)$ is the computational cost for each iteration. The multi-round estimator requires $t\big(T(n,p)+O(rnp)+O(mnp)\big) = O\big(tT(n,p)+tmnp\big)$ where $t$ is the number of rounds. In fact, we normally have $mnp = o(T(n,p))$ when $n$ and $p$ are large. For instance, the standard interior-point method for solving the LP in quantile regression of size $n$ by $p$ requires the computational cost of $O\big(n^{1.25}p^3\log n\big)$, cf. Portnoy and Koenker (1997). Therefore, the computational costs for the one-shot and multi-round estimators are $O\big(mT(n,p)\big)$ and $O\big(tT(n,p)\big)$, respectively. Our simulation results suggest that the multi-round estimators achieve desirable convergence rates in few rounds even when the number of machines is large (see Section C.6 in the Appendix). This implies the multi-round estima-

tor may require less computational time than the one-shot estimator when the number of machines is large, which is indeed consistent with our simulation results, cf. Section C.7 in the Appendix. For the communication cost, the one-shot estimator requires one round pass of $O(m)$ while the multi-round estimator requires $O(tms)$ in $t$ rounds where $s$ is the sparsity parameter.

Now we show the asymptotic normality of our multi-round estimator. We will first prove an asymptotic linear representation for the general one-step estimator (3.3), from which the asymptotic normality of our estimator follows. To this end, we make the following assumption. Recall the Gram matrix $\mathbf{\Sigma} = \mathbb{E}[\boldsymbol{X}\boldsymbol{X}^T]$. Let $c_1$ and $C_1$ be positive constants independent of $n$; $K_n \to \infty$ and $V_n \to \infty$ be sequences of positive constants. Define $c_n := (K_n+V_n)\vee K_nV_n\vee 1$, $a_n := \max(p, n, \mathrm{e})$ and $\rho_n = \rho(n) := c_n\sqrt{s\log a_n/n}$.

**Assumption 8.** *(i)* $\mathbb{E}[X_1^4] + \mathbb{E}[v^4] \le C_1$, $\mathbb{E}[v^2] \ge c_1$ *and* $\mathbb{E}[|v|^3 \mid \boldsymbol{X}_{(-1)}] \le C_1$ *a.s.; (ii)* $K$ *is a symmetric continuously differentiable kernel function with compact support such that* $\|K'\|_\infty < \infty$*; (iii)* $\|\boldsymbol{X}\|_\infty \le K_n$ *and* $|v| \le V_n$ *a.s.; (iv) The density of* $\epsilon$*,* $f_\epsilon(t)$*, is twice differentiable such that* $f_\epsilon(0) \ge c_1$*,* $\|f_\epsilon\|_\infty \vee \|f_\epsilon'\|_\infty \vee \|f_\epsilon''\|_\infty \le C_1$*; (v)* $\bar{\phi}_{\max}(\ell_n s, \mathbf{\Sigma}) \le C_1$ *for some* $\ell_n \to \infty$ *such that* $K_n^2\ell_n s\log p\log^3(K_n^2\ell_n s\log p) = o(n)$.

Condition (i) imposes moment conditions on the structural errors and covariates in order to apply arguments in the proof of Theorem 2 in Belloni et al. (2015b). Condition (ii) assumes mild conditions on the kernel function $K(t)$. Condition (iii) is assumed to be consistent with the theory in Section 3.3.1. Condition (iv) puts standard smoothness conditions on the density $f_\epsilon(t)$ that are common in the quantile regression literature Koenker (2005). Condition (v) is used to control the sparse eigenvalues of the empirical Gram matrix as discussed after Assumption 6.

To simplify the presentation of our inference results, we will impose a convergence rate condition on the estimators of $\boldsymbol{\beta}_0$ and $\boldsymbol{\theta}_0$. We will call $\hat{\boldsymbol{\beta}}$ *sparsely* $\rho_n$*-consistent* for $\boldsymbol{\beta}_0$ if it satisfies $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| \leq C_1\rho_n$ and $\|\hat{\boldsymbol{\beta}}\|_0 \leq C_1 s$ with probability $1 - o(1)$ for some constant $C_1$ independent of $n$. Based on the results of the last section, it is not difficult to see that the multi-round estimators of $\boldsymbol{\beta}_0$ and $\boldsymbol{\theta}_0$ are sparsely $\rho_N$-consistent under mild conditions.

Now we prove a linear representation of the one-step estimator. For the ease of presentation, we renumber the $N$ data on the $m$ machines as $\{(\boldsymbol{X}_j, y_j)\}_{j=1}^N$ when we present our theoretical results in the following. Define $\varphi(t) := 1/2 - I(t \leq 0)$ and recall that $h$ is the bandwidth parameter in the algorithm.

**Proposition 3.** *Suppose that that we observe $N$ i.i.d. data, $(\boldsymbol{X}_j, y_j)$ $(1 \leq j \leq N)$ and that Assumption 8 holds. Assume further the following conditions (i) $\hat{\boldsymbol{\beta}}_0$ and $\hat{\boldsymbol{\theta}}_0$ are sparsely $\rho_N$-consistent; (ii) $h = O\big(\rho_N^{1/2}\big)$; (iii) $c_n^6\sqrt{s \log a_n/N} = o(1)$. Then the one-step estimator (3.3) has the following representation,*

$$\check{\alpha}_1 = \alpha_1 + [f_\epsilon(0)\mathbb{E}(v^2)]^{-1} \cdot \mathbb{E}_N[\varphi(U_j - 1/2)v_j] + O_P\big(c_n(s \log a_n/N)^{3/4}\big),$$

*where $U_1, \cdots, U_N$ are i.i.d. uniform random variables on $(0, 1)$ independent from $v_1, \cdots, v_N$.*

The following corollary is immediate from the proposition, which shows the asymptotic normality of $\check{\alpha}_1^{\mathrm{ma}}$.

**Corollary 3.** *Suppose that we are running Algorithm 6. Under the assumption of Proposition 3, the one-step estimator of Algorithm 6 is asymptotically normal, $\sigma_n^{-1}\sqrt{N}(\check{\alpha}_1^{\mathrm{ma}} - \alpha_1) \xrightarrow{d} N(0, 1)$, where $\sigma_n^2 = \big\{4f_\epsilon(0)^2\mathbb{E}[v^2]\big\}^{-1}$.*

Proposition 3 also leads to an upper bound on the number of machines ($m$) for the one-shot estimator $\check{\alpha}_1^{\mathrm{oa}}$ to be asymptotically normal. Specifically, we ap-

ply Proposition 3 to the local one-step estimators from individual machines by replacing $N$ with the local sample size $n$. Then we can easily see that the one-shot estimator $\check{\alpha}_1^{\mathrm{oa}}$ would require

$$m = o\Big(\frac{n^{1/2}}{c_n^2(s\log a_n)^{3/2}}\Big)$$

to guarantee the averaged linear term dominates the remainder term in the expansion so that the asymptotic normality holds. To find the constraint on $m$ for the multi-round estimator $\check{\alpha}_1^{\mathrm{ma}}$, we shall appeal to Theorems 6 and 7 that guarantee the sparse $\rho_N$-consistency condition in Corollary 3. From Theorems 6 and 7, we see that the multi-round estimator $\check{\alpha}_1^{\mathrm{ma}}$ requires

$$m = O\big(\frac{n}{s^3K_n^2\log n}\big) \gg o\Big(\frac{n^{1/2}}{c_n^2(s\log a_n)^{3/2}}\Big)$$

for large $n$ if, for instance, $K_n$ and $V_n$ are constants independent of $n$. Thus, by using the multi-round estimation, we can allow much more machines to do inference than the naive one-shot approach, which is supported by our numerical experiments in Section 3.4.

**Simultaneous inference via pivotal bootstrap**

In this section, we consider simultaneous inference for multiple coefficients based on a new pivotal bootstrap for our multi-round one-step estimator. For a better presentation, we first redefine the notations we used before. Recall that we assume we have $m$ machines and each machine stores $n$ data, $(\boldsymbol{X}_{ij}, y_{ij})$ $(1 \le i \le m, 1 \le j \le n)$, that are generated from the following model,

$$y = \boldsymbol{X}^T\boldsymbol{\beta}_0 + \epsilon,\ \epsilon \perp\!\!\!\perp \boldsymbol{X} \quad \text{and} \quad \mathbb{P}(\epsilon \le 0) = 1/2, \tag{3.12}$$

where $\boldsymbol{\beta}_0 = (\alpha_1, \alpha_2, \cdots, \alpha_{p_1}, \beta_{0p_1+1}, \cdots, \beta_{0p})^T$ and we assume $\|\boldsymbol{\beta}_0\|_0 \le s$. We will construct simultaneous confidence intervals for the coefficients $\boldsymbol{\alpha} =$

$(\alpha_1, \cdots, \alpha_{p_1})^T$. Additionally, we assume the following auxiliary model for these $p_1$ covariates, for $1 \le k \le p_1$,

$$X_k = \boldsymbol{X}_{(-k)}^T \boldsymbol{\theta}_{(k)} + v_{(k)}, \quad \mathbb{E}[v_{(k)} | \boldsymbol{X}_{(-k)}] = 0, \tag{3.13}$$

where $\boldsymbol{X}_{(-k)} = (X_1, X_2, \cdots, X_{k-1}, X_{k+1}, \cdots, X_p)^T$ and we assume $\max_{1 \le k \le p_1} \|\boldsymbol{\theta}_{(k)}\|_0 \le s$. We will run Algorithm 6 to compute the one-step estimator $\check{\boldsymbol{\alpha}} = (\check{\alpha}_1, \cdots, \check{\alpha}_{p_1})^T$ of $\boldsymbol{\alpha}$ (note the multi-round quantile regression in the step 1 of Algorithm 6 is common among the computation of $\check{\alpha}_k$ for different $1 \le k \le p_1$). For the ease of the presentation of our theoretical results, we renumber the $N$ data on the $m$ machines as $\{(\boldsymbol{X}_j, y_j)\}_{j=1}^N$.

To provide valid simultaneous confidence intervals, we will first show that, under regularity conditions, the distribution of $\sqrt{N}(\check{\boldsymbol{\alpha}} - \boldsymbol{\alpha})$ can be approximated by a $p_1$-dimensional Gaussian vector uniformly over rectangles in $\mathbb{R}^{p_1}$ even when $p_1$ is far larger than $n$. However, the approximating normal vector has unknown variance which makes it infeasible for a direct use in practice. To deal with such difficulty, we propose a novel pivotal bootstrap to further approximate the sampling distribution.

Now, we prove our normal approximation result. To this end, we will first show a uniform linear representation of $(\check{\boldsymbol{\alpha}} - \boldsymbol{\alpha})$ which generalizes Proposition 3. We will need the following assumption which is a "uniform version" of Assumption 8. Recall Gram matrix $\boldsymbol{\Sigma} = \mathbb{E}[\boldsymbol{X}\boldsymbol{X}^T]$.

**Assumption 9.** *(i)* $\mathbb{E}[X_k^4] + \mathbb{E}[v_{(k)}^4] \le C_1$, $\mathbb{E}[v_{(k)}^2] \ge c_1$ *and* $\mathbb{E}[|v_{(k)}|^3 \mid \boldsymbol{X}_{(-k)}] \le C_1$ *a.s. uniformly in* $1 \le k \le p_1$*; (ii)* $\|\boldsymbol{X}\|_\infty \le K_n$ *and* $\max_{1 \le k \le p_1} |v_{(k)}| \le V_n$ *a.s.; (iii)* $\bar{\phi}_{\max}(\ell_n s, \boldsymbol{\Sigma}) \le C_1$ *for some* $\ell_n \to \infty$ *such that* $K_n^2 \ell_n s \log p \log^3(K_n^2 \ell_n s \log p) = o(N)$*; (iv) Conditions (ii) and (iv) in Assumption 8 hold.*

We also need a uniform version of sparsely $\rho_n$-consistency. Recall $\rho_n = \rho(n) := c_n\sqrt{s\log a_n/n}$ with $c_n := (K_n+V_n)\vee K_nV_n\vee 1$ and $a_n := \max(p,n,\mathrm{e})$. For a class of $p_1$ estimators $\{\hat{\boldsymbol{\gamma}}_1,\cdots,\hat{\boldsymbol{\gamma}}_{p_1}\}$ for $\{\boldsymbol{\gamma}_1,\cdots,\boldsymbol{\gamma}_{p_1}\}$, we will call such a class of estimators *uniformly sparsely $\rho_n$-consistent* if $\|\hat{\boldsymbol{\gamma}}_k-\boldsymbol{\gamma}_k\|\le C_1\rho_n$ and $\|\hat{\boldsymbol{\gamma}}_k\|_0\le C_1 s$ with probability $1-o(1)$ for some constant $C_1$ independent of $n$ uniformly in $1\le k\le p_1$. Now, we state the uniform linear representation of $\check{\boldsymbol{\alpha}}$ as the following proposition.

**Proposition 4.** *If Assumption 9 holds and assume further: (i) the class of estimators $\{\hat{\boldsymbol{\beta}}_0,\hat{\boldsymbol{\theta}}_{(1)},\cdots,\hat{\boldsymbol{\theta}}_{(p_1)}\}$ is uniformly sparsely $\rho_N$-consistent; (ii) $h = O\big(\rho_N^{1/2}\big)$; (iii) $c_n^6\sqrt{s\log a_n/N} = o(1)$. Then the following representation holds uniformly in $1\le k\le p_1$,*

$$\check{\alpha}_k = \alpha_k + [f_\epsilon(0)\mathbb{E}(v^2)]^{-1}\cdot\mathbb{E}_N[\varphi(U_j-1/2)v_{(k)j}] + O_P\big(c_n(s\log a_n/N)^{3/4}\big), \quad (3.14)$$

*where $U_1,\cdots,U_N$ are i.i.d. uniform random variables on $(0,1)$ independent from $v_{(k)1},\cdots,v_{(k)N}$ for any $1\le k\le p_1$.*

**Remark 11.** Condition (i) can be satisfied by our multi-round quantile regression and linear regression estimators. Specifically, Condition (i) will be satisfied if the high-dimensional quantile regression estimator $\hat{\boldsymbol{\beta}}_0$ is sparsely $\rho_N$-consistent and the collection of the heteroscedastic lasso estimators for the $p_1$ auxiliary models $\{\hat{\boldsymbol{\theta}}_{(1)},\cdots,\hat{\boldsymbol{\theta}}_{(p_1)}\}$ is uniformly sparsely $\rho_N$-consistent. The former can be achieved by our multi-round quantile regression, cf. Theorem 6. For the latter, inspection of the proof of Theorem 7 shows that the conclusions of Theorem 7 hold uniformly over $\hat{\boldsymbol{\theta}}_{(k)}$ $(1\le k\le p_1)$ if the conditions in Assumption 6 are strengthened to the corresponding "uniform versions" in Assumption 9. Therefore, the heteroscedastic lasso estimators $\{\hat{\boldsymbol{\theta}}_{(1)},\cdots,\hat{\boldsymbol{\theta}}_{(p_1)}\}$ is uniformly sparsely $\rho_N$-consistent.

The uniform linear representation (3.14) will be essential to show our following high-dimensional normal approximation.

First, we prepare some notations. Recall $\varphi(t) := 1/2 - I\{t \leq 0\}$. We introduce the influence function $\psi(x, y) := [f_\epsilon(0)\mathbb{E}(y^2)]^{-1}\varphi(x - 1/2)y$ and define $\Psi_j := \big(\psi(U_j, v_{(1)j}), \psi(U_j, v_{(2)j}), \cdots, \psi(U_j, v_{(p_1)j})\big)^T$. Define the normalization matrix $\boldsymbol{A} := 2f_\epsilon(0)\text{diag}\{\mathbb{E}[v_{(1)}^2]^{1/2}, \cdots, \mathbb{E}[v_{(p_1)}^2]^{1/2}\} := \text{diag}\{A_1, \cdots, A_{p_1}\}$. The following theorem presents the normal approximation result.

**Theorem 8** (Normal approximation)**.** *Suppose the conditions of Proposition 4 hold and we further assume following growth conditions*

$$\frac{V_n^2 \log^7(p_1 N)}{N} \to 0 \quad \textit{and} \quad \frac{c_n^4 (s \log a_n)^3 \log^2 p_1}{N} \to 0. \tag{3.15}$$

*Then we have*

$$\sup_{\boldsymbol{b} \in \mathbb{R}^{p_1}} \big|\mathbb{P}\big(\sqrt{N}\boldsymbol{A}(\check{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) \leq \boldsymbol{b}\big) - \mathbb{P}\big(\boldsymbol{AG} \leq \boldsymbol{b}\big)\big| \to 0,$$

*where* $\boldsymbol{G} \sim N(0, \boldsymbol{\Sigma}_1)$ *with* $\boldsymbol{\Sigma}_1 := \mathbb{E}[\Psi_j \Psi_j^T]$.

We emphasize that Condition (3.15) allows $p_1$ to be much larger than $N$, i.e., $p_1 \gg N$. The proof of Theorem 8 builds on the uniform asymptotic linear representation developed in Proposition 4 coupled with the high dimensional Gaussian approximation techniques developed in Chernozhukov et al. (2014, 2017a). Theorem 8 implies that the distribution of $\sqrt{N}\boldsymbol{A}(\check{\boldsymbol{\alpha}} - \boldsymbol{\alpha})$ can be approximated by the distribution of $\boldsymbol{AG}$ uniformly over the rectangles. However, the approximating normal vector is infeasible due to its unknown covariance matrix. To overcome this difficulty, we will use a new pivotal bootstrap to approximate the sampling distribution.

The proof of Theorem 8 shows that the distribution of $\boldsymbol{G}$ comes from approximating the distribution of

$$\frac{1}{\sqrt{N}}\sum_{j=1}^{N}\psi(U_j, v_{(j)}), \tag{3.16}$$

for $1 \le j \le p_1$. Importantly, (3.16) is "pivotal" in the sense that its distribution is completely known up to some estimable nuisance parameters given the data $\{(\boldsymbol{X}_j, y_j)\}_{j=1}^N$ since $U_1, \ldots, U_N$ are independent $U(0,1)$ random variables. The baseline idea of the pivotal bootstrap is to simulate the distribution of (3.16) (given the data) to estimate the distribution of $\boldsymbol{G}$ by generating $U(0,1)$ random variables while plugging in estimators of unknown nuisance parameters in (3.16).

Define the following bootstrap quantities, for $1 \le j \le N$ and $1 \le k \le p_1$,

$$\hat{\psi}(U_j, \hat{v}_{(k)j}) := [\hat{f}_\epsilon(0)\mathbb{E}_N(\hat{v}_{(k)j}^2)]^{-1}\varphi(U_j - 1/2)\hat{v}_{(k)j},$$

$$\hat{\Psi}_j := (\hat{\psi}(U_j, \hat{v}_{(1)j}), \hat{\psi}(U_j, \hat{v}_{(2)j}), \cdots, \hat{\psi}(U_j, \hat{v}_{(p_1)j}))^T,$$

$$\hat{\boldsymbol{A}} := 2\hat{f}_\epsilon(0)\mathrm{diag}\{\mathbb{E}_N[\hat{v}_{(1)j}^2]^{1/2}, \cdots, \mathbb{E}_N[\hat{v}_{(p_1)j}^2]^{1/2}\} := \mathrm{diag}\{\hat{A}_1, \cdots, \hat{A}_{p_1}\}.$$

The pivotal bootstrap reads as follows. Generate $U_1, \ldots, U_N \sim U(0,1)$ i.i.d. that are independent of the data $\mathcal{D}_n := (\boldsymbol{X}_j, y_j)_{j=1}^N$. We denote the conditional probability $\mathbb{P}(\cdot \mid \mathcal{D}_n)$ and conditional expectation $\mathbb{E}[\cdot \mid \mathcal{D}_n]$ by $\mathbb{P}_U(\cdot)$ and $\mathbb{E}_U[\cdot]$, respectively. Define $\hat{\boldsymbol{\Sigma}}_1 := N^{-1}\sum_{j=1}^N \mathbb{E}_U[\hat{\Psi}_j\hat{\Psi}_j^T]$. Then, we shall estimate the distribution of $\boldsymbol{AG}$ (or $N^{-1/2}\sum_{j=1}^N \boldsymbol{A}\Psi_j$) by the conditional distribution of $N^{-1/2}\sum_{i=1}^N \hat{\boldsymbol{A}}\hat{\Psi}_j$ given the data $\mathcal{D}_n$. The conditional distribution can be simulated with arbitrary precision. The following theorem establishes the consistency of the pivotal bootstrap over the rectangles.

**Theorem 9** (Pivotal bootstrap consistency)**.** *Suppose the conditions of Proposition 4 hold and we further assume following growth conditions*

$$\frac{(V_n + \rho_N)^2\log^7(p_1 N)}{N} \to 0 \quad \textit{and} \quad \left(\rho_N \vee c_n^{-1/4}N^{-3/8}(s\log a_n)^{-1/8}\right)\log^2 p_1 \to 0.$$

*Then we have*

$$\sup_{\boldsymbol{b}\in\mathbb{R}^{p_1}} \left|\mathbb{P}_U\big(N^{-1/2}\textstyle\sum_{j=1}^N \hat{\boldsymbol{A}}\hat{\Psi}_j \le \boldsymbol{b}\big) - \mathbb{P}\big(\boldsymbol{AG}\le \boldsymbol{b}\big)\right| \xrightarrow{P} 0.$$

The proof of Theorem 9 is nontrivial since the pivotal bootstrap differs from the nonparametric or multiplier bootstraps that have been analyzed in the literature in the high-dimensional setup. The proof consists of two steps. First, since $\hat{\Psi}_1,\ldots,\hat{\Psi}_N$ are independent with mean zero conditionally on the data $\mathcal{D}_n$, we apply the high dimensional CLT conditionally on $\mathcal{D}_n$ to approximate the conditional distribution of $N^{-1/2}\sum_{i=1}^N \hat{\boldsymbol{A}}\hat{\Psi}_i$ by the conditional Gaussian distribution $N(0,\hat{\boldsymbol{A}}\hat{\boldsymbol{\Sigma}}_1\hat{\boldsymbol{A}}^T)$. Second, we apply Gaussian comparison techniques to show the difference between $N(0,\hat{\boldsymbol{A}}\hat{\boldsymbol{\Sigma}}_1\hat{\boldsymbol{A}}^T)$ and the sampling distribution $\boldsymbol{AG}\sim N(0,\boldsymbol{A}\boldsymbol{\Sigma}_1\boldsymbol{A}^T)$ is asymptotically negligible.

Based on the proof of Theorem 9, we can show that the conclusion of Theorem 8 continues to hold even if the normalization matrix $\boldsymbol{A}$ is replaced by its estimate $\hat{\boldsymbol{A}}$.

**Theorem 10.** *Suppose conditions in the Theorem 8 holds and we further assume*

$$\big(\rho_N \vee c_n^{-1/4}N^{-3/8}(s\log a_n)^{-1/8}\big)\log p_1 \to 0.$$

*Then we have*

$$\sup_{\boldsymbol{b}\in\mathbb{R}^{p_1}} \left|\mathbb{P}\big(\sqrt{N}\hat{\boldsymbol{A}}(\check{\boldsymbol{\alpha}}-\boldsymbol{\alpha})\le \boldsymbol{b}\big) - \mathbb{P}\big(\boldsymbol{AG}\le \boldsymbol{b}\big)\right| \to 0.$$

Now, based on the Theorems 8 and 9, if we denote, for $0<\alpha<1$,

$$\hat{q}_{1-\alpha} = \text{conditional } (1-\alpha)\text{-quantile of } \max_{1\le k\le p_1}\left|N^{-1/2}\textstyle\sum_{j=1}^N \hat{A}_k\hat{\psi}(U_j,\hat{v}_{(k)j})\right|,$$

we can show that the data-dependent rectangle (interval when $p_1=1$)

$$\prod_{k=1}^{p_1}\left[\check{\alpha}_k \pm \frac{\hat{q}_{1-\alpha}}{\sqrt{N}\hat{A}_k}\right] \tag{3.17}$$

contains the vector $\boldsymbol{\alpha}$ with probability approaching $1-\alpha$. Formally, the coverage guarantee of the preceding confidence rectangle follows from

$$\mathbb{P}\left(\max_{1\leq k\leq p_1}\left|\sqrt{N}\hat{A}_k(\check{\alpha}_k-\alpha_k)\right|\leq \hat{q}_{1-\alpha}\right)\to 1-\alpha. \tag{3.18}$$

The latter (3.18) follows from Theorems 8 and 9 coupled with Lemma 1 in Zhang et al. (2021) which we restate as Lemma 1 in the Appendix for completeness.

We further illustrate the multi-round estimation of multiple coefficients and the pivotal bootstrap in a distributed setting in Algorithm 8. Algorithm 8 takes the advantage of the parallel computing in both estimation and bootstrap steps and minimizes the communication cost by only passing the processed quantities from the local machines back to the master machine according to the form of the one-step estimator.

## 3.4 Numerical examples

### 3.4.1 Implementation details

In our simulation, we use the biwight kernel and the default bandwidth selected by the R function *density()* for the density estimation of multi-round estimators while we directly plug in the the true density value for the one-shot estimator. Besides, we use post-$\ell_1$-penalized median regression and post-lasso regression for the one-shot procedure. For our multi-round estimation procedures, we use $\ell_1$-penalized median regression and regular lasso regression to compute the initial estimate.

In the implementation of our multi-round quantile and linear regression

**Algorithm 8:** Pivotal bootstrap for simultaneous confidence intervals

**Input**: Data: $\{(\boldsymbol{X}_{ij}, y_{ij})\}$ $(1 \le i \le m, 1 \le j \le n)$; Median score function: $\varphi(t) = 1/2 - I(t \le 0)$; Kernel function: $K(t)$; Bandwidth: $h$.

**Estimation**:

**Step 1**. Run multi-rounded adaptively weighted median regression and obtains the final estimators $\hat{\boldsymbol{\beta}}_0$ of $\boldsymbol{\beta}_0$; For $1 \le k \le p_1$, run multi-rounded heteroscedastic Lasso of $X_k$ on $\boldsymbol{X}_{(-k)}$ and obtains the final estimators $\hat{\boldsymbol{\theta}}_{(k)}$ of $\boldsymbol{\theta}_{(k)}$. Machine 1 broadcasts $\hat{\boldsymbol{\beta}}_0$ and $\hat{\boldsymbol{\theta}}_{(k)}$ $(1 \le k \le p_1)$ to the rest of machines and keeps the initial biased estimator $\hat{\alpha}_1, \cdots, \hat{\alpha}_{p_1}$ obtained from $\hat{\boldsymbol{\beta}}_0$

**Step 2**. For machine $i$ $(1 \le i \le m)$, it computes $\mathbb{E}_n[h^{-1}K(\hat{\epsilon}_{ij}/h)]$, $\mathbb{E}_n[\varphi(\hat{\epsilon}_{ij})\hat{v}_{ij(k)}]$ and $\mathbb{E}_n[\hat{v}^2_{ij(k)}]$ for $1 \le k \le p_1$ where $\hat{\epsilon}_{ij} = y_{ij} - \boldsymbol{X}_{ij}^T\hat{\boldsymbol{\beta}}_0$ and $\hat{v}_{ij(k)} = X_{ijk} - \boldsymbol{X}^T_{ij(-k)}\hat{\boldsymbol{\theta}}_{(k)}$; Machine $i$ passes these quantities back to Machine 1.

**Step 3**. Machine 1 forms the one-step estimator of $\alpha_k$ $(1 \le k \le p_1)$ as following,

$$\check{\alpha}_k = \hat{\alpha}_k + \big[m^{-2}\textstyle\sum_{i=1}^m \mathbb{E}_n[h^{-1}K(\hat{\epsilon}_{ij}/h)] \cdot \sum_{i=1}^m \mathbb{E}_n[\hat{v}^2_{ij(k)}]\big]^{-1} \cdot \big[m^{-1}\textstyle\sum_{i=1}^m \mathbb{E}_n[\varphi(\hat{\epsilon}_{ij})\hat{v}_{ij(k)}]\big].$$

**Pivotal bootstrap**:
**for** $l = 1, \cdots, T$ **do**
  **for** $i = 1, \cdots, m$ **do**
    Machine $i$ generates $U_{ijl} \sim U(0,1)$ i.i.d. $(1 \le j \le n)$; Then machine $i$ formulate $\big(\sum_{j=1}^n [\varphi(U_{ijl} - 1/2)\hat{v}_{ij(k)}]\big)_{k=1}^{p_1}$ and pass it to Machine 1.
  **end**
  Machine 1 computes

$$q_l := \max_{1 \le k \le p_1} \Big| 2N^{-1/2}[m^{-1}\textstyle\sum_{i=1}^m \mathbb{E}_n\{\hat{v}^2_{ij(k)}\}]^{-1/2} \cdot \sum_{i=1}^m \sum_{j=1}^n [\varphi(U_{ijl} - 1/2)\hat{v}_{ij(k)}] \Big|$$

**end**
Machine 1 takes $\hat{q}_{1-\alpha} :=$ $(1-\alpha)$-quantile of $\{q_l\}_{l=1}^T$ and compute the simultaneous confidence intervals of level $(1-\alpha)$ according to (3.17).

estimators, we have to select several tuning parameters. For the multi-round quantile regression estimator, we select $\omega_{t+1,k}$ $(1 \le k \le p)$ and $\lambda_{t+1}$ in the $(t+1)$-round. In the simulation, we take $\omega_{t+1,k} = \hat{\beta}^{-1}_{t,k}$ where $\hat{\beta}_t$ is the estimator from the $t$-round. We select $\lambda_{t+1}$ based on a "cross-validation" type procedure. Specifi-

cally, we pick a grid for $\lambda_{t+1}$ and compute the estimators over the grid. Then we choose the optimal $\lambda_{t+1}$ corresponding to the estimator that minimizes the total quantile loss on Machine $2$ to Machine $m$. We also select the tunning parameter $\lambda_{t+1}$ in the multi-round lasso estimator using this approach by minimizing the least square loss on Machine $2$ to Machine $m$. For a fair comparison, we select the tunning parameters in the one-shot procedure by the same approach.

We refer to Remark 9 for more details on solving the optimization problems of the multi-round quantile regression. To solve the optimization problem in (3.9) for the multi-round lasso, we use the accelerated proximal gradient algorithm which is implemented in R package *apg* Vert (2015).

In the following simulation study, we fix the number of rounds to be $5$ when computing the multi-round estimators. Sensitivity analysis of the number of rounds is provided in Section C.6 of the Appendix. We also report the computational time comparison between the multi-round one-step estimator and the one-shot estimator in Section C.7 of the Appendix.

All the results are obtained in the R environment with 28 Intel Xeon processors and 240 Gbytes RAM over Red Hat OpenStack Platform.

### 3.4.2 Pointwise confidence intervals

In this section, we consider confidence intervals for a single coefficient. We generate data from the following models similar to those considered in Belloni et al. (2015b),

$$y = X_1\alpha_1 + \boldsymbol{X}_{(-1)}^T\boldsymbol{\theta}_0 + \epsilon, \quad X_1 = \boldsymbol{X}_{(-1)}^T\boldsymbol{\theta}_0 + v,$$

where $\alpha_1 = 1$ and $\boldsymbol{X}_{(-1)} = (1, \boldsymbol{Z})$ with $\boldsymbol{Z}$ containing the rest of regressors. We generate $\boldsymbol{Z} \sim N(0, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}$ has entries $\Sigma_{ij} = 0.5^{|i-j|}$. We take $\boldsymbol{\theta}_0$ to be $400$ dimensional, i.e., $p = 401$ and

$$\boldsymbol{\theta}_0^T = (0.2, 1, 0, 1/2^2, \ldots, 1/8^2, 0, \ldots, 0).$$

We generate $v \sim N(0, 1)$ and consider following three generation distributions for $\epsilon$,

- Normal: $\epsilon \overset{i.i.d.}{\sim} N(0, 1)$.
- $\mathrm{T}_2$: $\epsilon \overset{i.i.d.}{\sim} \mathrm{T}(2)$ (t-distribution with degrees of freedom 2).
- Exponential: $\epsilon \overset{i.i.d.}{\sim} \exp(1)$.

We note that the variance of the $\mathrm{T}_2$ distribution is infinite. We consider three different total sample sizes $N = 6000, 9000, 12000$ and three different local sample sizes $n = 300, 500, 1000$. Similar settings are considered in Jordan et al. (2019) for the high-dimensional regression. For each of these nine situations, we compare the performance of our pivotal bootstrap confidence interval based on the multi-round estimators (multi-round CIs) and the confidence interval based on the one-shot estimator (one-shot CIs) and its asymptotic normality. We take the nominal level of the confidence intervals to be $95\%$. The results are based on 400 repetitions in each case. We report the empirical coverage probabilities (Coverage), average confidence interval lengths (Length) and the standard deviations of lengths of the resulting confidence intervals (SD) in Table 3.1 to Table 3.3.

From the tables, the multi-round CIs achieve close to nominal level coverage probabilities in all the scenarios while the one-shot CIs fail in most of them. We found that the main reason for low coverage probabilities of the one-shot CIs is

due to its accumulating bias. This can be seen form the decrease in the coverage probabilities as the total sample size $N$ increases when the local sample size $n$ is fixed. Meanwhile, multi-round CIs maintain satisfying coverage probabilities when we increase the number of machines which supports our theories. Besides, the lengths of the multi-round CIs are very close to the asymptotic normality based one-shot CIs both of which decrease with the increase of the total sample size. This justifies our normal approximation results and the consistency of the pivotal bootstrap and also suggests sufficiently fast convergence of the bootstrap approximations.

Table 3.1: Pointwise confidence intervals for Normal model.

| $n$ | | 300 | | | 500 | | | 1000 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $N$ | | 6000 | 9000 | 12000 | 6000 | 9000 | 12000 | 6000 | 9000 | 12000 |
| MR | Coverage(%) | 94.5 | 93.5 | 95.25 | 96 | 95.25 | 95.75 | 93 | 94.5 | 96.25 |
| | Length | 0.065 | 0.052 | 0.045 | 0.064 | 0.052 | 0.045 | 0.064 | 0.053 | 0.045 |
| | SD($\times 10^{-3}$) | 3.4 | 2.5 | 2.1 | 3.2 | 2.6 | 2.0 | 3.3 | 2.4 | 2.2 |
| OS | Coverage(%) | 76.75 | 62.5 | 61 | 86 | 80.25 | 77.5 | 92.25 | 89 | 88 |
| | Length | 0.065 | 0.053 | 0.046 | 0.065 | 0.053 | 0.046 | 0.064 | 0.052 | 0.045 |
| | SD($\times 10^{-3}$) | 0.66 | 0.42 | 0.32 | 0.63 | 0.39 | 0.29 | 0.58 | 0.38 | 0.30 |

Table 3.2: Pointwise confidence intervals for $\mathrm{T}_2$ model.

| $n$ | | 300 | | | 500 | | | 1000 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $N$ | | 6000 | 9000 | 12000 | 6000 | 9000 | 12000 | 6000 | 9000 | 12000 |
| MR | Coverage(%) | 95.25 | 95.75 | 95.5 | 93.75 | 97.25 | 94 | 95.75 | 95 | 96.25 |
| | Length | 0.074 | 0.061 | 0.052 | 0.074 | 0.060 | 0.052 | 0.074 | 0.060 | 0.052 |
| | SD($\times 10^{-3}$) | 3.7 | 2.9 | 2.4 | 3.8 | 2.9 | 2.4 | 3.6 | 2.8 | 2.4 |
| OS | Coverage(%) | 67 | 62.75 | 49.75 | 85.75 | 78 | 75.75 | 91.75 | 88.5 | 89.5 |
| | Length | 0.074 | 0.060 | 0.052 | 0.073 | 0.060 | 0.051 | 0.072 | 0.059 | 0.051 |
| | SD($\times 10^{-3}$) | 0.66 | 0.45 | 0.35 | 0.70 | 0.47 | 0.33 | 0.72 | 0.45 | 0.34 |

Table 3.3: Pointwise confidence intervals for Exponential model.

| $n$ | | 300 | | | 500 | | | 1000 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $N$ | | 6000 | 9000 | 12000 | 6000 | 9000 | 12000 | 6000 | 9000 | 12000 |
| MR | Coverage(%) | 94.25 | 92.25 | 93.5 | 96 | 94.5 | 95.5 | 94.25 | 92.75 | 94 |
| | Length | 0.050 | 0.040 | 0.035 | 0.050 | 0.041 | 0.035 | 0.050 | 0.041 | 0.035 |
| | SD($\times 10^{-3}$) | 2.7 | 2.2 | 1.8 | 2.6 | 2.2 | 1.8 | 2.6 | 2.1 | 1.7 |
| OS | Coverage(%) | 77.5 | 70.5 | 61.25 | 89.75 | 84 | 83.25 | 94 | 93.5 | 90.25 |
| | Length | 0.052 | 0.042 | 0.037 | 0.052 | 0.042 | 0.036 | 0.051 | 0.042 | 0.036 |
| | SD($\times 10^{-3}$) | 0.49 | 0.35 | 0.24 | 0.50 | 0.34 | 0.25 | 0.49 | 0.33 | 0.25 |

### 3.4.3 Simultaneous confidence intervals

In this section, we consider simultaneous confidence intervals for two coefficients. We generate data from the following models,

$$y = X_1\alpha_1 + X_2\alpha_2 + \boldsymbol{X}_{(-12)}^T\boldsymbol{\theta}_0 + \epsilon, \quad X_1 = \boldsymbol{X}_{(-12)}^T\boldsymbol{\theta}_0 + v_1, \quad X_2 = \boldsymbol{X}_{(-12)}^T\boldsymbol{\theta}_0 + v_2,$$

where $\alpha_1 = 2$, $\alpha_2 = 1$ and $\boldsymbol{X}_{(-12)} = (1, \boldsymbol{Z})$ with $\boldsymbol{Z}$ containing the rest of regressors. We take $\boldsymbol{\theta}_0$ as in the pointwise case and generate $\boldsymbol{Z}$ in the same way as there. We generate $v_1,\, v_2 \sim N(0,1)$ and consider the three generation distributions for $\epsilon$ as in the pointwise case. We consider the same settings for the local sample size ($n$) and the total sample size ($N$) as in Section 3.4.2. Given the unsatisfactory performance and extensive computational cost of the one-shot estimation in the pointwise case, we only consider the multi-round one-step estimator here. Similarly, we consider $95\%$ simultaneous confidence intervals. We report the coverage probabilities (Coverage), average length of simultaneous bootstrap confidence intervals (Length) and their standard deviations (SD). The results are based on 400 repetitions in each case and presented in Table 3.4.

From the table, the multi-round CIs again achieve satisfying coverage probabilities in all the considered scenarios which provides strong support for our theoretical results. The length of multi-round CIs decrease with the increase of

the total sample size $N$ as predicted by the theory. Under the same scenario, the length of the simultaneous multi-round CI is larger than the pointwise case. This is due to the fact that the simultaneous confidence intervals are accounting for more parameters.

Table 3.4: Simultaneous confidence intervals for the multi-round one-step estimator.

| $n$ | | 300 | | | 500 | | | 1000 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $N$ | | 6000 | 9000 | 12000 | 6000 | 9000 | 12000 | 6000 | 9000 | 12000 |
| Nor | Coverage(%) | 93.25 | 95.75 | 94.75 | 93.75 | 94.25 | 94.5 | 94 | 96.25 | 93.75 |
| | Length | 0.074 | 0.060 | 0.052 | 0.073 | 0.060 | 0.052 | 0.074 | 0.060 | 0.052 |
| | SD($\times 10^{-3}$) | 3.2 | 2.4 | 1.9 | 3.2 | 2.4 | 2.1 | 3.2 | 2.3 | 2.0 |
| $T_2$ | Coverage | 94.75% | 95.5% | 96% | 94% | 94.25% | 94.5% | 94.25% | 95% | 95% |
| | Length | 0.085 | 0.069 | 0.060 | 0.084 | 0.068 | 0.046 | 0.085 | 0.069 | 0.059 |
| | SD($\times 10^{-3}$) | 4.0 | 3.1 | 2.4 | 3.7 | 2.7 | 2.2 | 3.6 | 2.7 | 2.4 |
| Exp | Coverage(%) | 92.25 | 94.5 | 94.75 | 93.25 | 94 | 94.25 | 93.75 | 93.25 | 94.5 |
| | Length | 0.057 | 0.046 | 0.040 | 0.057 | 0.046 | 0.040 | 0.058 | 0.047 | 0.041 |
| | SD($\times 10^{-3}$) | 2.8 | 2.3 | 1.9 | 2.7 | 2.0 | 1.7 | 2.6 | 2.0 | 1.7 |

## 3.5 Summary

In this paper, we develop a multi-round aggregated one-step estimator for the distributed high-dimensional LAD regression based on new multi-round quantile regression and heteroscedastic linear regression estimators. We derive convergence rates and sparsity properties of the new multi-round estimators and show that our multi-round one-step estimator requires less restrictive sample complexity than the straightforward one-shot aggregation. We also propose a new pivotal bootstrap for simultaneous inference and establish its validity allowing for the number of the inferred coefficients exceeding the sample size. Both the new multi-round one-step estimator and the pivotal bootstrap are scalable to the distributed setting. The numerical results provide strong sup-

port of our theoretical results. Several interesting extensions remain, including the extension to the heteroscedastic quantile regression model. Under such a model, we need to modify the one-step estimator and the bootstrap method while taking both the heteroscedasticity and practical constraints into consideration, which requires much new and delicate analysis. These are beyond the scope of the current paper and left for future research.

APPENDIX A

**APPENDIX OF CHAPTER 1**

## A.1 Extra Notation

For a matrix $A \in \mathbb{R}^{p\times p}$, $\lambda_{\max}(A)$ denotes the largest eigenvalue of matrix $A$ and $||A||$ is the Frobenius norm of $A$. For two positive definite matrices, $B$ and $C$, we write $B > C$ if and only if $B - C$ is positive definite (p.d.). In particular, $B$ is p.d. if and only if $B > 0$.

## A.2 Proof of Theorem 1

To show the consistency of $\hat{\beta}_n$, we will need the following lemma.

**Lemma 2.** *For any $\beta \in \mathcal{B}$, if $\zeta(\beta) := E\left[\{b'(X^T\beta) - Y\}X\right]$ is finite and the following condition holds*

$$\sum_{i=1}^{n} E\left[\frac{\{b'(X_i^T\beta) - Y_i\}^2}{\pi_i} x_{ij}^2\right] = o(n^2 r), \tag{A.1}$$

*then we have $\Psi_n^*(\beta) - \zeta(\beta) \xrightarrow{p} 0$ for any $\beta \in \mathcal{B}$.*

*Proof.* Observe that

$$\begin{aligned} E\Psi_n^*(\beta) &= E\left[\frac{1}{r}\sum_{i=1}^{r} \frac{b'(X_i^{*T}\beta) - Y_i^*}{n\pi_i^*} \cdot X_i^*\right] \\ &= E\left[E\left\{\frac{1}{r}\sum_{i=1}^{r} \frac{b'(X_i^{*T}\beta) - Y_i^*}{n\pi_i^*} \cdot X_i^* | (X_i, Y_i)_{i=1}^n\right\}\right] \\ &= \zeta. \end{aligned}$$

We will use Chebyshev's inequality to show convergence in probability. We denote the $j$-th element in the vector $(\Psi_n^*)$ by

$$(\Psi_n^*)_j = \frac{1}{r}\sum_{i=1}^{r} \frac{b'(X_i^{*T}\beta) - Y_i^*}{n\pi_i^*} \cdot x_{ij}^*,$$

and the $j$-th coordinate of $\zeta$ as $\zeta_j$.

By using Chebyshev's inequality, it suffices to show $E[(\Psi_n^*)_j - \zeta_j]^2 = o(1)$, where

$$E[(\Psi_n^*)_j - \zeta_j]^2 = E\left[E\left\{[(\Psi_n^*)_j - \zeta_j]^2 \,|(X_i, Y_i)_{i=1}^n\right\}\right].$$

For this expectation

$$\begin{aligned} E\left[E\left\{[(\Psi_n^*)_j - \zeta_j]^2 \,|(X_i, Y_i)_{i=1}^n\right\}\right] &= E\left[\frac{1}{r^2}\cdot\sum_{i=1}^{r} E_*\left\{\frac{b'(X_i^{*T}\beta) - Y_i^*}{n\pi_i^*}\cdot x_{ij}^* - \zeta_j\right\}^2\right] \\ &= E\left[\frac{1}{r}\sum_{i=1}^{n}\pi_i\left\{\frac{b'(X_i^T\beta) - Y_i}{n\pi_i}\cdot x_{ij} - \zeta_j\right\}^2\right] \\ &= \frac{1}{rn^2}\sum_{i=1}^{n} E\left[\frac{\left\{b'(X_i^T\beta) - Y_i\right\}^2}{\pi_i}\cdot x_{ij}^2\right] - \frac{1}{r}\zeta_j^2, \end{aligned}$$

where the first equality is based on the fact that after conditioning on the $n$ data points, the $r$ repeating sampling steps should be independent and distributionally identical in each step. And we use $E_*$ to denote expectation with respect to sampling randomness. Hence, we have

$$E[(\Psi_n^*)_j - \zeta_j]^2 = \frac{1}{rn^2}\sum_{i=1}^{n} E\left[\frac{\left(b'\left(X_i^T\beta\right) - Y_i\right)^2}{\pi_i}\cdot x_{ij}^2\right] - \frac{1}{r}\zeta_j^2 = o(1).$$

The second equality is due to the assumption (A.1). □

We now show the consistency of $\hat{\beta}_n$.

We will verify the conditions in Theorem 5.9 in van der Vaart (2000) and apply the theorem to prove the consistency of $\hat{\beta}_n$.

First of all, for any $\beta_1$ and $\beta_2$ in $\mathcal{B}$,

$$\begin{aligned}
&|[\Psi_n^*(\beta_1) - \zeta(\beta_1)] - [\Psi_n^*(\beta_2) - \zeta(\beta_2)]| \\
=&\left\| \left\{ \frac{1}{r} \sum_{i=1}^{r} \frac{b''(X_i^{*T}\tilde{\beta}_1)}{n\pi_i^*} \cdot X_i^* X_i^{*T} - E\left[ b''(X^T\tilde{\beta}_2) X X^T \right] \right\} \cdot (\beta_1 - \beta_2) \right\| \\
\leq&\left\| \frac{1}{r} \sum_{i=1}^{r} \frac{b''(X_i^{*T}\tilde{\beta}_1)}{n\pi_i^*} \cdot X_i^* X_i^{*T} - E\left[ b''(X^T\tilde{\beta}_2) X X^T \right] \right\| \cdot \|\beta_1 - \beta_2\| \\
:=&L_n \cdot ||\beta_1 - \beta_2||.
\end{aligned}$$

The first step is due to mean value theorem with $\tilde{\beta}_1$ and $\tilde{\beta}_2$ lying on the segment between $\beta_1$ and $\beta_2$.

We now show $L_n = O_p(1)$. By assumption (i), it suffices to show $\frac{1}{r}\sum_{i=1}^{r} \frac{X_i^* X_i^{*T}}{n\pi_i^*} = O_p(1)$. This is true because we have

$$\begin{aligned}
&E\left[ \frac{1}{r} \sum_{i=1}^{r} \frac{X_i^* X_i^{*T}}{n\pi_i^*} \right] \\
=&E\left[ E\left\{ \frac{1}{r} \sum_{i=1}^{r} \frac{X_i^* X_i^{*T}}{n\pi_i^*} \middle| (X_i, Y_i)_{i=1}^n \right\} \right] \\
=&EXX^T,
\end{aligned}$$

and it follows from Markov inequality. Now we apply Lemma 2.9 in Newey and McFadden (1986) to conclude $\Psi_n^*(\beta) - \zeta(\beta)$ is stochastic equicontinuous. Again, by Theorem 21.9 in Davidson (1994), Lemma 2 and stochastic equicontinuity imply

$$\sup_{\beta \in \mathcal{B}} \|\Psi_n^*(\beta) - \zeta(\beta)\| \xrightarrow{p} 0.$$

This uniform convergence condition together with condition (iv) in the theorem yield the desired conclusion by applying Theorem 5.9 in van der Vaart (2000).

□

## A.3 Proof of Theorem 2

In this section, we will establish the asymptotic normality of $\hat{\beta}_n$. Let us start will some auxiliary lemmas.

### A.3.1 Proof of Lemma 1

We first prove following lemma.

**Lemma 3.** *Assume that* $\Phi = E\left\{b''(X^T\beta_0)XX^T\right\}$ *is finite and non-singular. Further assume for* $1 \leq k, j \leq p$,

$$\sum_{i=1}^{n} E\left\{\frac{b''(X_i^T\beta_0)^2}{\pi_i}(x_{ik}x_{ij})^2\right\} = o(n^2 r). \tag{A.2}$$

*Then, we have* $\dot{\Psi}_n^*(\beta_0) = \frac{1}{r}\sum_{i=1}^{r} \frac{b''(X_i^{*T}\beta_0)}{n\pi_i^*} \cdot X_i^* X_i^{*T} \xrightarrow{p} \Phi$.

*Proof.* First we note

$$\begin{aligned} E\dot{\Psi}_n^*(\beta_0) &= E\left[\frac{1}{r}\sum_{i=1}^{r} \frac{b''(X_i^{*T}\beta_0)}{n\pi_i^*} \cdot X_i^* X_i^{*T}\right] \\ &= E\left[E\left\{\frac{1}{r}\sum_{i=1}^{r} \frac{b''(X_i^{*T}\beta_0)}{n\pi_i^*} \cdot X_i^* X_i^{*T}\middle|(X_i, Y_i)_{i=1}^n\right\}\right] = \Phi. \end{aligned}$$

To show the convergence in probability, we use Chebyshev's inequality. Consider each element in the matrix

$$\begin{aligned} (\dot{\Psi}_n^*)_{kj} &= \frac{1}{r}\sum_{i=1}^{r} \frac{b''(X_i^{*T}\beta_0)}{n\pi_i^*} \cdot x_{ik}^* x_{ij}^*, \\ \Phi_{kj} &= E\left[b''(X^T\beta_0)x_k x_j\right]. \end{aligned}$$

By using Chebyshev's inequality, it suffices to show $E\left[(\dot{\Psi}_n^*)_{kj} - \Phi_{kj}\right]^2 = o(1)$.

$$E\left[(\dot{\Psi}_n^*)_{kj} - \Phi_{kj}\right]^2 = E\left[E\left\{\left[(\dot{\Psi}_n^*)_{kj} - \Phi_{kj}\right]^2 \middle| (X_i, Y_i)_{i=1}^n\right\}\right].$$

For this expectation

$$\begin{aligned}
E\left[E\left\{\left[(\dot{\Psi}_n^*)_{kj} - \Phi_{kj}\right]^2 \middle| (X_i, Y_i)_{i=1}^n\right\}\right] &= E\left[\frac{1}{r^2}\cdot\sum_{i=1}^{r} E_*\left[\frac{b''(X_i^{*T}\beta_0)}{n\pi_i^*}\cdot x_{ik}^* x_{ij}^* - \Phi_{kj}\right]^2\right] \\
&= E\left[\frac{1}{r}\sum_{i=1}^{n}\pi_i\left[\frac{b''(X_i^T\beta_0)}{n\pi_i}\cdot x_{ik}x_{ij} - \Phi_{kj}\right]^2\right] \\
&= \frac{1}{rn^2}\sum_{i=1}^{n} E\left[\frac{b''(X_i^T\beta_0)^2}{\pi_i}\cdot x_{ik}^2 x_{ij}^2\right] - \frac{1}{r}\Phi_{kj}^2,
\end{aligned}$$

where the first equality is based on the fact that after conditioning on the $n$ data points, the $r$ repeating sampling steps should be independent and distributionally identical in each step, and we use $E_*$ to denote expectation with respect to sampling randomness. Hence, we have

$$E\left[(\dot{\Psi}_n^*)_{kj} - \Phi_{kj}\right]^2 = \frac{1}{rn^2}\sum_{i=1}^{n} E\left[\frac{b''(X_i^T\beta_0)^2}{\pi_i}\cdot x_{ik}^2 x_{ij}^2\right] - \frac{1}{r}\Phi_{kj}^2 = o(1).$$

The second equality is due to the assumption (A.2). □

Now we prove Lemma 1.

By Taylor's Theorem:

$$0 = \Psi_n^*(\hat{\beta}_n) = \Psi_n^*(\beta_0) + \dot{\Psi}_n^*(\beta_0)(\hat{\beta}_n - \beta_0) + \frac{1}{2}(\hat{\beta}_n - \beta_0)^T\ddot{\Psi}_n^*(\tilde{\beta}_n)(\hat{\beta}_n - \beta_0),$$

where $\tilde{\beta}_n$ is on the line segment between $\beta_0$ and $\hat{\beta}_n$. $\ddot{\Psi}_n^*$ is a k-vector of $(k \times k)$ matrices.

We now show that $\left\|\ddot{\Psi}_n^*(\tilde{\beta}_n)\right\| = O_p(1)$. By assumption (iv)

$$\begin{aligned}
\left\|\ddot{\Psi}_n^*(\tilde{\beta}_n)\right\| &= \left\|\frac{1}{r}\sum_{i=1}^{r}\frac{1}{n\pi_i^*}\cdot\ddot{\psi}_{\tilde{\beta}_n}(X_i^*)\right\| \\
&\leq \frac{1}{r}\sum_{i=1}^{r}\frac{1}{n\pi_i^*}\cdot\left\|\ddot{\psi}(X_i^*)\right\| = O_p(1).
\end{aligned}$$

The last equality is because of the fact

$$E\left[\frac{1}{r}\sum_{i=1}^{r}\frac{1}{n\pi_i^*}\cdot\left\|\ddot{\psi}(X_i^*)\right\|\right]=E\left\|\ddot{\psi}(X)\right\|=\text{const}$$

and application of Markov inequality.

Therefore,

$$0=\Psi_n^*(\beta_0)+(\Phi+o_p(1))\left(\hat{\beta}_n-\beta_0\right)+O_p\left(\left\|\hat{\beta}_n-\beta_0\right\|^2\right).$$

This implies the conclusion

$$\Psi_n^*(\beta_0)=-\Phi(\hat{\beta}_n-\beta_0)+o_p\left(\left\|\hat{\beta}_n-\beta_0\right\|\right).$$

□

### A.3.2 Multivariate martingale CLT

Now, we prove a multivariate extension of the martingale central limit theorem stated in Ohlsson (1989) Theorem A.1, which will be appropriate for our with replacement sampling setting.

**Lemma 4** (Multivariate version of martingale CLT)**.** *For $k=1,2,3,\ldots$, let $\{\xi_{ki};i=1,2,\ldots,N_k\}$ be a martingale difference sequence in $\mathbb{R}^p$ relative to the filtration $\{\mathcal{F}_{ki};i=0,1,\ldots,N_k\}$ and let $Y_k\in\mathbb{R}^p$ be an $\mathcal{F}_{k0}$-measurable random vector. Set $S_k=\sum_{i=1}^{N_k}\xi_{ki}$. Assume the following conditions.*

*(i)* $\lim_{k\to\infty}\sum_{i=1}^{N_K}E\left[\|\xi_{ki}\|^4\right]=0$

*(ii)* $\lim_{k\to\infty}E\left[\left\|\sum_{i=1}^{N_k}E\left[\xi_{ki}\xi_{ki}^T|\mathcal{F}_{k,i-1}\right]-B_k\right\|^2\right]=0$ *for some sequence of positive definite matrices $\{B_k\}_{k=1}^{\infty}$ with $\sup_k\lambda_{max}(B_k)<\infty$ i.e. the largest eigenvalue is uniformly bounded.*

*(iii) For some probability distribution $L_0$, $*$ denotes convolution and $L(\cdot)$ denotes the law of random variables:*

$$L(Y_k) * N(0, B_k) \xrightarrow{d} L_0.$$

*Then we have*

$$L(Y_k + S_k) \xrightarrow{d} L_0.$$

*Proof.* We use Cramer-Wold device to deduce it from the univariate case. For any $a \in \mathbb{R}^p$, by Cramer-Wold device, it suffices to show

$$L(a^T Y_k + a^T S_k) \xrightarrow{d} a^T L_0.$$

We check the conditions of Theorem A.1 in Ohlsson (1989).

1. $\sum_{i=1}^{N_k} E\left[\left(a^T \xi_{ki}\right)^4\right] \leq \sum_{i=1}^{N_k} ||a||^4 \cdot E\left[||\xi_{ki}||^4\right] = ||a||^4 \sum_{i=1}^{N_k} E[||\xi_{ki}||^4] \longrightarrow 0$ The inequality is due to Cauchy-Schwarz inequality.

2.
$$\begin{aligned}
&E\left[\sum_{i=1}^{N_k} E\left[a^T \xi_{ki} \xi_{ki}^T a | \mathcal{F}_{k,i-1}\right] - a^T B_k a\right]^2 \\
=&E\left[a^T \left\{\sum_{i=1}^{N_k} E[\xi_{ki}\xi_{ki}|\mathcal{F}_{k,i-1}] - B_k\right\} a\right]^2 \\
\lesssim&E\left[\left\|\sum_{i=1}^{N_k} E[\xi_{ki}\xi_{ki}|\mathcal{F}_{k,i-1}] - B_k\right\| \cdot ||a||^2\right]^2 \longrightarrow 0 \text{ as } k \to \infty.
\end{aligned}$$

3.
$$\begin{aligned}
\phi_{a^T Y_k} \cdot \phi_{N(0, a^T B_k a)} =& E\left[e^{ita^T Y_k}\right] \cdot e^{-\frac{1}{2}(a^T B_k a)t^2} \\
=& E\left[e^{i\xi^T Y_k}\right] \cdot e^{-\frac{1}{2}\xi^T B_k \xi} \qquad \text{where } \xi = at \\
\longrightarrow& \phi_{L_0}(at) \equiv \phi_{a^T L_0}(t).
\end{aligned}$$

Here we use $\phi_*(t)$to denote the characteristic function. Hence

$$L(a^T Y_k) * N(0, a^T B_k a) \xrightarrow{d} a^T L_0.$$

From above verification, we use Theorem A.1 in Ohlsson (1989) to obtain

$$L(a^T Y_k + a^T S_k) \xrightarrow{d} a^T L_0.$$

And by Cramer-Wold device, this finishes the proof. □

### A.3.3 More Auxiliary Results

**Lemma 5.** *$\{M_i\}_{i=1}^r$ is a martingale difference sequence relative to the filtration $\{\mathcal{F}_{n,i}\}_{i=1}^r$.*

*Proof.* The $\mathcal{F}_{n,i}$-measurability follows from the definition of $M_i$ and the definition of the filtration $\{\mathcal{F}_{n,i}\}_{i=1}^r$. And we also have

$$E[M_i|\mathcal{F}_{n,i-1}] = E_{*_i}\left[\frac{b'(X_i^{*T}\beta) - Y_i^*}{rn\pi_i^*} \cdot X_i^*\right] - \frac{1}{rn}\sum_{j=1}^{n}\left(b'(X_j^T\beta_0) - Y_j\right) \cdot X_j = 0.$$

Combine these two results, we finish the proof. □

With Lemma 4, we could easily get the following result.

**Corollary 1.** $V(T) = V(M) + V(Q)$.

**Lemma 6.** $\sup_n \lambda_{max}(B_n) \leq 1$.

*Proof.* Since $B_n$ is symmetric, it suffices to show for any $n$, $I - B_n$ is positive definite.

$$\begin{aligned} I - B_n &= V(T)^{-\frac{1}{2}}\left(V(T) - V(M)\right)V(T)^{-\frac{1}{2}} \\ &= V(T)^{-\frac{1}{2}}V(Q)V(T)^{-\frac{1}{2}}. \end{aligned}$$

Therefore, $I - B_n$ is congruent to matrix $V(Q)$ which is positive definite. Hence $I - B_n$ is also positive definite and this finishes the proof. □

**Lemma 7** (Asymptotic normality of $\Psi_n^*(\beta_0)$)**.** *Assume the following conditions*

*(i)* $\lim_{n\to\infty} \sum_{i=1}^{r} E\left[||\xi_{ni}||^4\right] = 0.$

*(ii)* $\lim_{n\to\infty} E\left[\left\|\sum_{i=1}^{r} E[\xi_{ni}\xi_{ni}^T|\mathcal{F}_{n,i-1}] - B_n\right\|^2\right] = 0.$

*Then we will have*

$$V(T)^{-\frac{1}{2}} \cdot T \xrightarrow{d} N(0, I).$$

*Proof.* We verify the conditions in Lemma 4 with

$$\xi_{ki} = \xi_{ni}, \qquad Y_k = V(T)^{-\frac{1}{2}} \cdot Q,$$
$$B_k = B_n, \qquad L_0 \sim N(0, I).$$

By Lemma 5, conditions (i) and (ii), we can easily see the first two conditions of Lemma 4 are satisfied. It suffices to show the third condition in Lemma 4 holds. We first note the following conclusion

$$V(Q)^{-\frac{1}{2}} Q \xrightarrow{d} N(0, I).$$

This because $Q$ is a sum of i.i.d mean zero random variables, $(b'(X_j^T\beta_0) - Y_j) \cdot X_j$, which have finite variance and a simple application of central limit theorem will give the above conclusion.

Now, we verify the third condition. For any $t \in \mathbb{R}^p$

$$
\begin{aligned}
&E[e^{it^TV(T)^{-\frac{1}{2}}Q}] \cdot e^{-\frac{1}{2}t^TV(T)^{-\frac{1}{2}}V(M)V(T)^{-\frac{1}{2}}t} \\
=&\left(Ee^{it^TV(T)^{-\frac{1}{2}}V(Q)V(T)^{-\frac{1}{2}}t} + o(1)\right) \cdot e^{-\frac{1}{2}t^TV(T)^{-\frac{1}{2}}V(M)V(T)^{-\frac{1}{2}}t} \\
=&Ee^{it^TV(T)^{-\frac{1}{2}}V(Q)V(T)^{-\frac{1}{2}}t} \cdot e^{-\frac{1}{2}t^TV(T)^{-\frac{1}{2}}V(M)V(T)^{-\frac{1}{2}}t} + o(1) \\
=&e^{-\frac{1}{2}t^Tt} + o(1).
\end{aligned}
$$

The first equality is due to Lemma 8 in the following. Therefore, we have verified the third condition in Lemma 4. And by that lemma we have

$$V(T)^{-\frac{1}{2}} \cdot Q + V(T)^{-\frac{1}{2}} \cdot M = V(T)^{-\frac{1}{2}}T \xrightarrow{d} N(0, I).$$

□

Now we state the following lemma that has been used in the proof of previous lemma.

**Lemma 8.** *Under conditions in Lemma 7 For any $t \in \mathbb{R}^p$,*

$$\left|E\left[e^{it^TV(T)^{-\frac{1}{2}}Q}\right] - E\left[e^{it^TV(T)^{-\frac{1}{2}}V(Q)^{\frac{1}{2}}A_0}\right]\right| \longrightarrow 0$$

*as $n \to \infty$ and $A_0 \sim N(0, I)$.*

*Proof.* Since $V(Q)^{-\frac{1}{2}}Q \xrightarrow{d} N(0, I)$, for any $\xi \in \mathbb{R}^p$,

$$\left|E\left[e^{i\xi^TV(Q)^{-\frac{1}{2}}Q}\right] - E\left[e^{i\xi^TA_0}\right]\right| \longrightarrow 0$$

as $n \to \infty$. And the convergence is uniform in any finite set of $\xi$. (see Chapter 6 of Chung (2001)). By setting $\xi = V(Q)^{\frac{1}{2}}V(T)^{-\frac{1}{2}}t^T$, to prove the lemma, it suffices to show

$$\sup_n ||\xi|| < \infty.$$

for any fixed $t$. Also we note that

$$||\xi|| \leq \lambda_{max}\left(V(Q)^{\frac{1}{2}}V(T)^{-\frac{1}{2}}\right)\cdot||t||.$$

Hence, it is enough to show $\lambda_{max}\left(V(Q)^{\frac{1}{2}}V(T)^{-\frac{1}{2}}\right)\leq 1$. For notation simplicity, we denote $A = V(Q)$ and $B = V(T)$ in the following proof of the lemma.Note the following equation holds

$$A^{\frac{1}{2}}B^{-\frac{1}{2}} = B^{\frac{1}{4}}\left(B^{-\frac{1}{4}}A^{\frac{1}{2}}B^{-\frac{1}{4}}\right)B^{-\frac{1}{4}}$$

That is $A^{\frac{1}{2}}B^{-\frac{1}{2}}$ is similar to $B^{-\frac{1}{4}}A^{\frac{1}{2}}B^{-\frac{1}{4}}$. Therefore, we only need to show $\lambda_{max}\left(B^{-\frac{1}{4}}A^{\frac{1}{2}}B^{-\frac{1}{4}}\right)\leq 1$. This is implied by the fact

$$I - B^{-\frac{1}{4}}A^{\frac{1}{2}}B^{-\frac{1}{4}} > 0.$$

In fact

$$I - B^{-\frac{1}{4}}A^{\frac{1}{2}}B^{-\frac{1}{4}} = B^{-\frac{1}{4}}\left(B^{\frac{1}{2}} - A^{\frac{1}{2}}\right)B^{-\frac{1}{4}}.$$

That is, $I - B^{-\frac{1}{4}}A^{\frac{1}{2}}B^{-\frac{1}{4}}$ is congruent to $B^{\frac{1}{2}} - A^{\frac{1}{2}}$. Therefore, it suffices to show $B^{\frac{1}{2}} - A^{\frac{1}{2}}$ is positive definite.

Note $B > A$. This is because $B - A = V(M) > 0$. Now we use Theorem 1.1 in Zhan (2004), then we will have $B^{\frac{1}{2}} - A^{\frac{1}{2}} > 0$ which finishes proof. □

Now we are able to prove our theorem 2 which shows the asymptotic normality of the sampling estimator $\hat{\beta}_n$.

### A.3.4 Proof of Theorem 2

*Proof.* By Lemma 1

$$\Phi(\hat{\beta}_n - \beta_0) + o_p\left(\left\|\hat{\beta}_n - \beta_0\right\|\right) = -\Psi_n^*(\beta_0).$$

Now we normalize both sides with $V(T)^{-\frac{1}{2}}$

$$V(T)^{-\frac{1}{2}}\Phi(\hat{\beta}_n - \beta_0) + o_p\left(\left\|V(T)^{-\frac{1}{2}}\hat{\beta}_n - \beta_0\right\|\right) = -V(T)^{-\frac{1}{2}}\Psi_n^*(\beta_0).$$

By Lemma 7

$$V(T)^{-\frac{1}{2}}\Phi(\hat{\beta}_n - \beta_0) \xrightarrow{d} N(0, I).$$

□

## A.4 Proof of Theorem 3

First of all, we condition on $X_1^n$(or consider $X_1^n$ is fixed). We now find out $V(T|X_1^n) = V(\Psi_n^*(\beta_0)|X_1^n)$. We have

$$V(T|X_1^n) = E_Y\left[V(T|X_1^n, Y_1^n)\right] + V_Y\left[E(T|X_1^n, Y_1^n)\right].$$

Here $E_Y$ means we take expectation w.r.t randomness of $Y$ after we conditioning on $X$. After some simple calculation, we could get

$$V_Y[E(T|X_1^n, Y_1^n)] = \frac{1}{n^2}\sum_{j=1} b''(X_j^T\beta_0)\cdot X_j X_j^T,$$
$$E_Y[V(T|X_1^n, Y_1^n)] = \frac{1}{n^2 r}\sum_{j=1}^{n} b''(X_j^T\beta_0)\cdot X_j X_j^T\cdot\left(\frac{1}{\pi_j} - 1\right).$$

Hence, we have

$$V(T|X_1^n) = \frac{1}{n^2}\sum_{j=1}^{n} b''(X_j^T\beta_0)\cdot X_j X_j^T\cdot\left(\frac{1}{r\pi_j} - \frac{1}{r} + 1\right).$$

We now minimize $tr(\Phi^{-1}V(T|X_1^n)\Phi^{-1})$

$$\begin{aligned}
tr(\Phi^{-1}V(T|X_1^n)\Phi^{-1}) &= \frac{1}{n^2}\sum_{j=1}^{n} tr\left(b''(X_j^T\beta_0)\cdot\Phi^{-1}X_jX_j^T\Phi^{-1}\cdot(\frac{1}{r\pi_j}-\frac{1}{r}+1)\right)\\
&= \frac{1}{rn^2}\sum_{j=1}^{n} tr\left(\frac{b''(X_j^T\beta_0)}{\pi_j}\cdot\Phi^{-1}X_jX_j^T\Phi^{-1}\right)+C\\
&= \frac{1}{rn^2}\sum_{j=1}^{n}\frac{b''(X_j^T\beta_0)}{\pi_j}\cdot\left\|\Phi^{-1}X_j\right\|^2+C\\
&= \frac{1}{rn^2}\sum_{j=1}^{n}\pi_j\sum_{j=1}^{n}\frac{b''(X_j^T\beta_0)}{\pi_j}\cdot\left\|\Phi^{-1}X_j\right\|^2+C\\
&\geq \frac{1}{rn^2}\left(\sum_{j=1}^{n}\sqrt{b''(X_j^T\beta_0)}\cdot\left\|\Phi^{-1}X_j\right\|\right)^2+C,
\end{aligned}$$

where in the last step we use Cauchy-Schwarz inequality and the equality holds iff $\pi_j \propto \sqrt{b''(X_j^T\beta_0)}\,\|\Phi^{-1}X_j\|$.

Now we consider $V(T)$ under random design.

$$V(T) = E\left[V(T|X_1^n)\right] + V\left[E(T|X_1^n)\right].$$

However, we could verify that in GLM

$$E(T|X_1^n) \equiv 0.$$

Therefore, we have $V(T) = E[V(T|X_1^n)]$. From this we have

$$\begin{aligned}
\{\pi_j^{opt}\}_{j=1}^n &= \arg\min_{\pi} tr(\Phi^{-1}V(T)\Phi^{-1})\\
&= \arg\min_{\pi} tr\left(E\left[\Phi^{-1}V(T|X_1^n)\Phi^{-1}\right]\right)\\
&= \arg\min_{\pi} E\left[tr\left(\Phi^{-1}V(T|X_1^n)\Phi^{-1}\right)\right]\\
&= \arg\min_{\pi} tr\left(\Phi^{-1}V(T|X_1^n)\Phi^{-1}\right).
\end{aligned}$$

## A.5 Additional Plots of Section 1.5

### A.5.1 Computational Time Plots for Logistic Regression

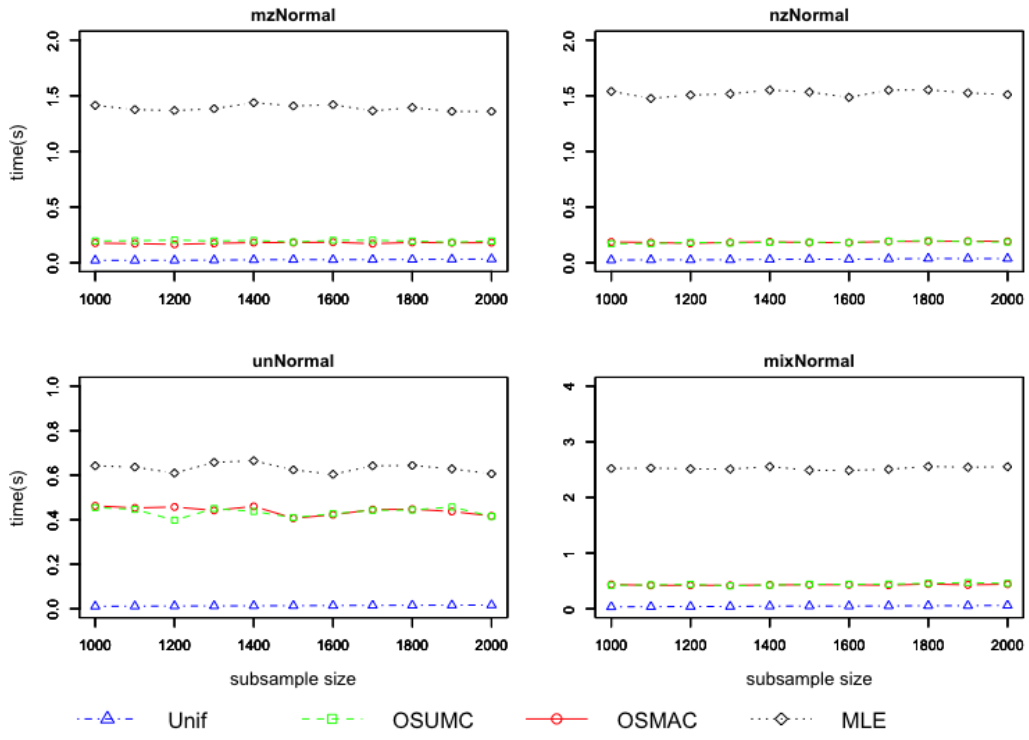
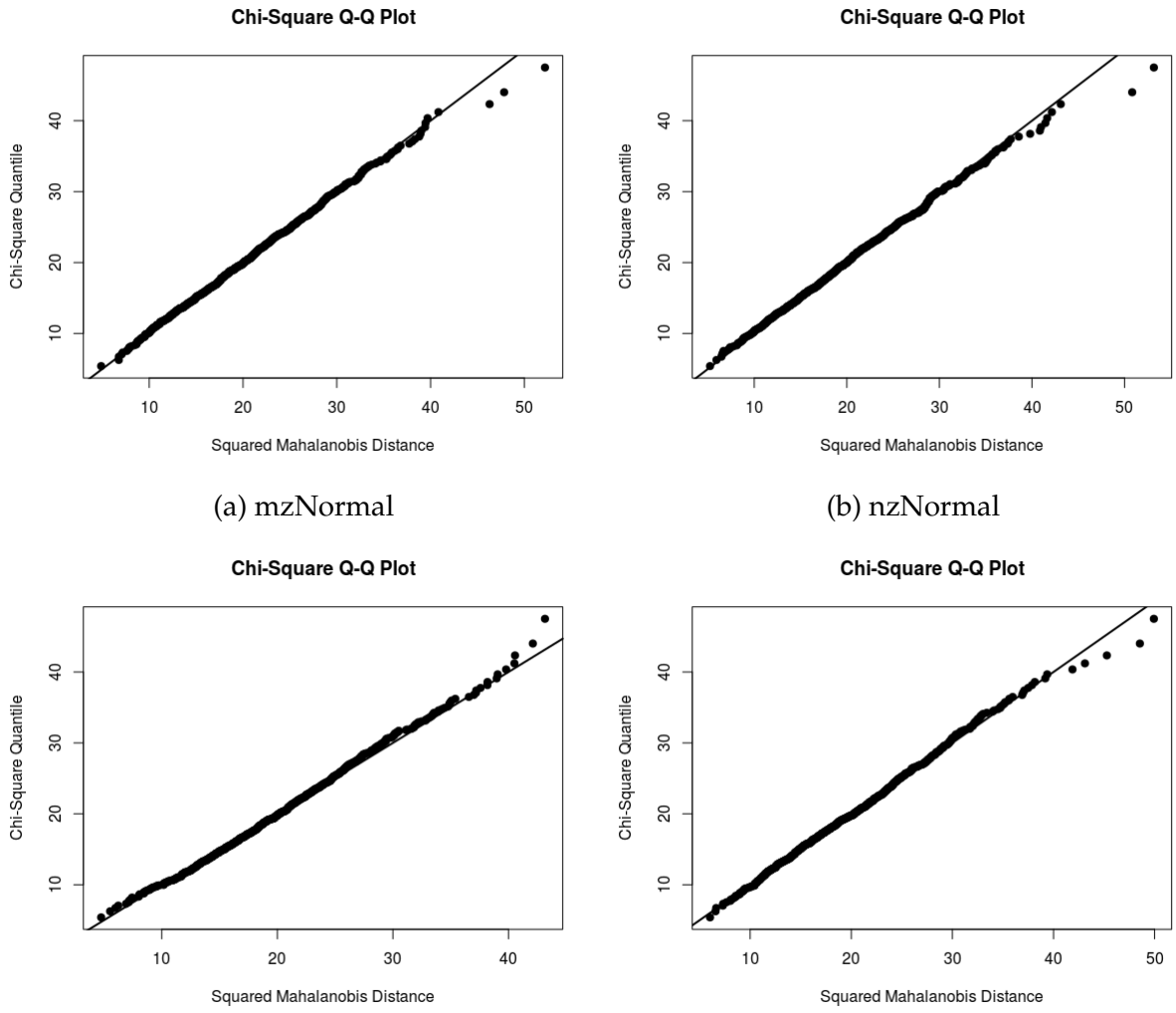



Figure A.1: Computational time for different subsample size $r$ under different design generation settings for logistic regression with $r_0 = 500$

Figure A.1 reveals that the computation time is not very sensitive to the subsample size for all the four methods. All of the three sampling methods outperform the full sample MLE. It is not surprising to see the uniform sampling always takes the least computation time, since it does not involve the computation of the sampling probability to compensate for the loss of efficiency. In most cases, OSUMC and OSMAC require significantly less computational time compared with full-sample MLE.

### A.5.2 Q-Q Plots for Logistic Regression

To see whether the asymptotic normality in our theory implies approximate finite-sample normality under the previous four different design generations in logistic regression model, we plot the chi-square Q-Q plot of the resultant estimator $\hat{\beta}_n$ for each considered setting. Here, we replace the approximated optimal sampling weight in Algorithm 2 with the oracle optimal weights to calculate the estimator $\hat{\beta}_n$, i.e., the true $\beta_0$ is used in the calculation of optimal sampling weights. Experiments are repeated $1000$ times under each setting and corresponding Q-Q plots are presented in the Figure A.2. Nearly all the points lie on the straight line in each plot, which is consistent with $\hat{\beta}_n$ being approximately normally distributed in the four considered design generation settings.




(a) mzNormal

(b) nzNormal

(c) unNormal

(d) mixNormal

Figure A.2: Chi-square Q-Q plots of $\hat{\beta}_n$ under different design generation settings for logistic regression with $r = 5000$.

### A.5.3 Computational Time Plots for Linear Regression

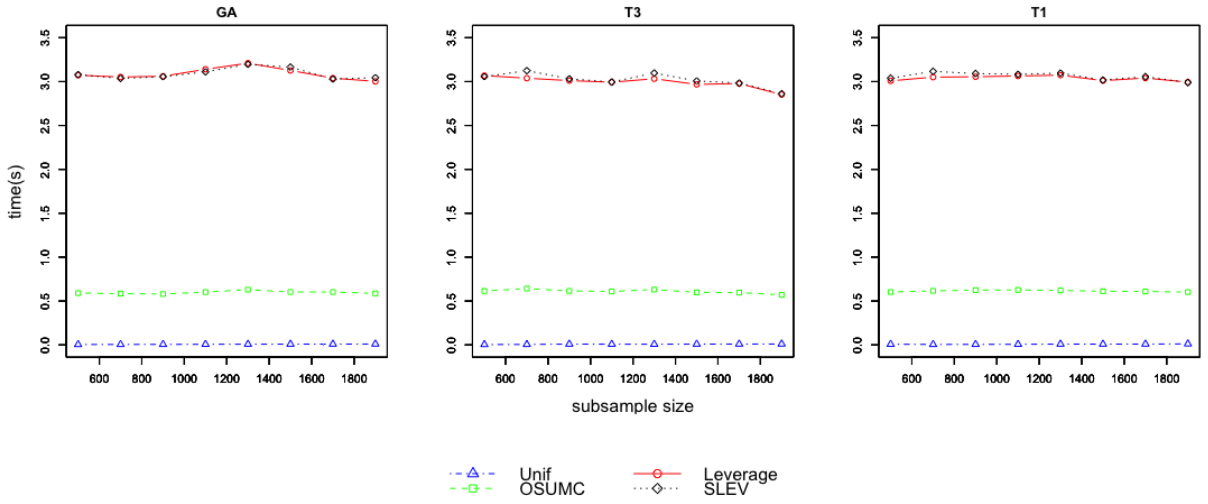


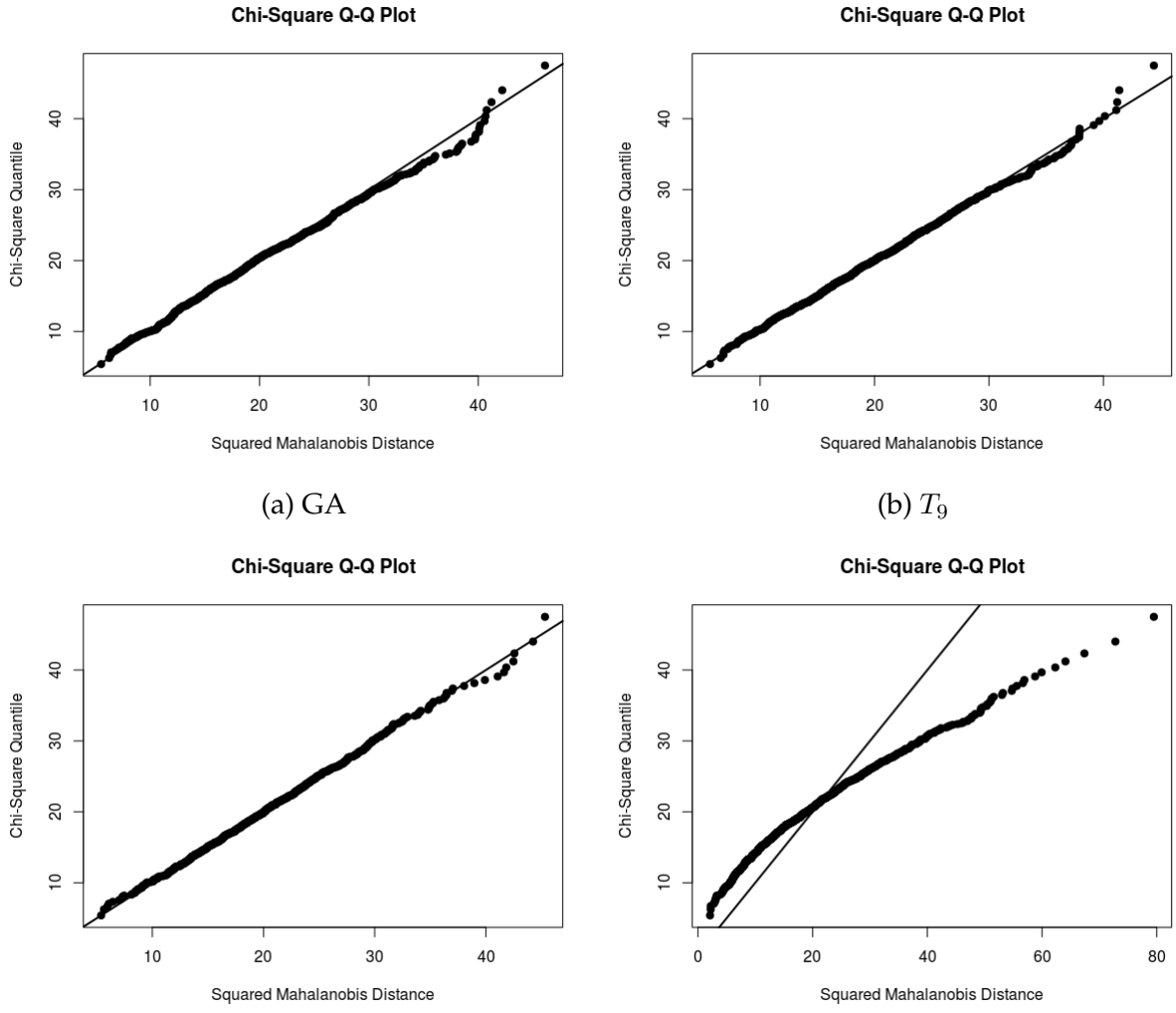

Figure A.3: Computational time plots for different subsample size $r$ under different design generation settings for linear regression

From Figure A.3, Again, the results show the insensitivity of the computational time to increasing subsample sizes. Our method requires the second smallest computing time, being inferior only to the uniform sampling. Both leverage related methods take more than double the computational time of our method due to the intensive computation of leveraging score of each data point.

### A.5.4 Q-Q Plots for Linear Regression

To explore further how sensitive the approximate finite-sample normality is to the moment condition of the design distribution in linear regression setting, we show chi-square Q-Q plots for several design generation distributions with different orders of moment. To be more specific, GA, $T_9$, $T_3$ and $T_1$ distributions

are considered. Experiments are repeated $1000$ times under each setting and results are presented in the following.




(a) GA (b) $T_9$

(c) $T_3$ (d) $T_1$

Figure A.4: Chi-square Q-Q plots of $\hat{\beta}_n$ under different design generation settings for linear regression with $r = 5000$.

As shown in Figure A.4, the resultant sampling estimator $\hat{\beta}_n$ is approximately normal in GA, $T_9$ and $T_3$ settings where we should note that the multivariate t-distribution with 3 degrees of freedom doesn't even have a third moment. This indicates that the normality of $\hat{\beta}_n$ in linear models holds under very

weak moment conditions for the design-generation distribution. One surprising fact is that OSUMC outperforms the other sampling methods in the $T_1$ setting, even though normality fails to hold.

### A.5.5 Computational Time Plot for Superconductivity Data Set

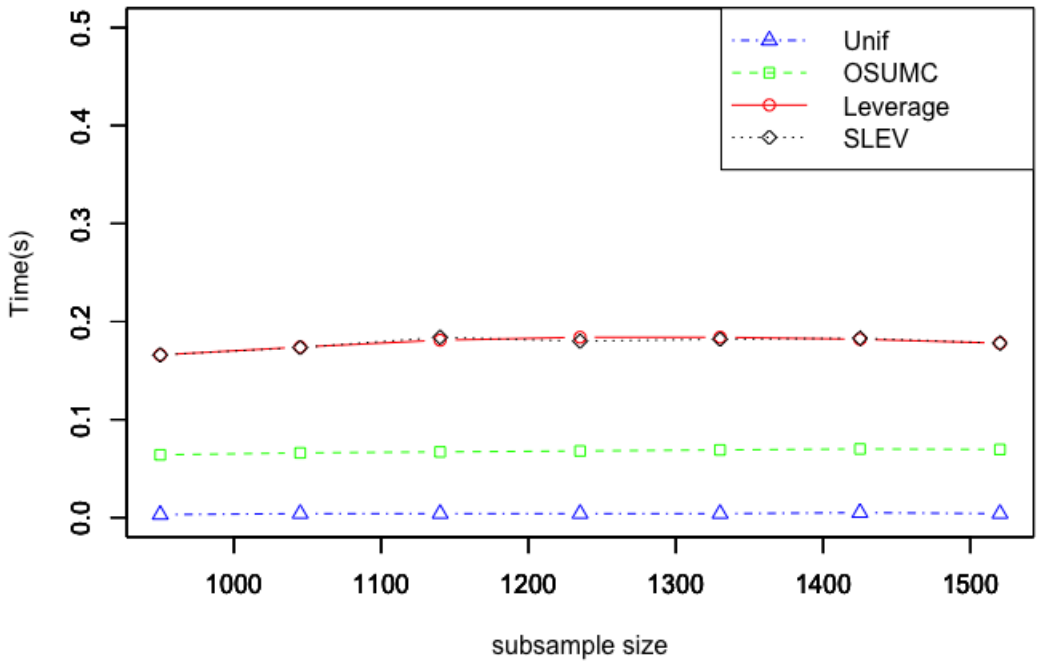



Figure A.5: Computational time plots for different subsample size

## A.6 Simulation for Poisson Regression

We generate datasets of size $n = 100,000$ from the following Poisson regression model,

$$Y \sim \text{Poisson}\left(\exp\{\mathrm{X}^{\mathrm{T}}\beta_0\}\right),$$

where $\beta_0$ is a $100$ dimensional vector with all entries $0.5$. We consider two different scenarios to generate $X$.

- **Case 1.** each covariate of $X$ follows independent uniform distribution over $[-0.5, 0.5]$.
- **Case 2.** First half of covariates of $X$ follow independent uniform distribution over $[-0.5, 0.5]$ while the other half follow independent uniform distribution over $[-1, 1]$.

In each case, we compare our optimal sampling procedure (OSUMC) with uniform sampling (Unif), and the benchmark full data MLE under different subsample sizes. In our procedure, $r_0$, the subsample size in the first step uniform sampling, equals $500$. For uniform sampling, we directly subsample $r$ points and calculate the subsample MLE. Again, we repeat the simulation $500$ times and report the empirical MSE and computational time in Figures A.6 and A.7, respectively.

From Figure A.6, OSUMC method uniformly dominates the uniform sampling method in both scenarios in terms of mean squared errors, which supports the A-optimality of OSUMC. For average computational time, our simulation reveals that the computation time is not very sensitive to the subsample size. In both cases, OSUMC requires significantly less computational time compared with full-sample MLE.

To see whether the asymptotic normality in our theory holds under the both design generation settings, we plot the chi-square Q-Q plot of the resultant estimator $\hat{\beta}_n$ for each considered setting and the results are presented in Figure A.8. Q-Q plots reveal that $\hat{\beta}_n$ is approximately normal with sufficiently large sample

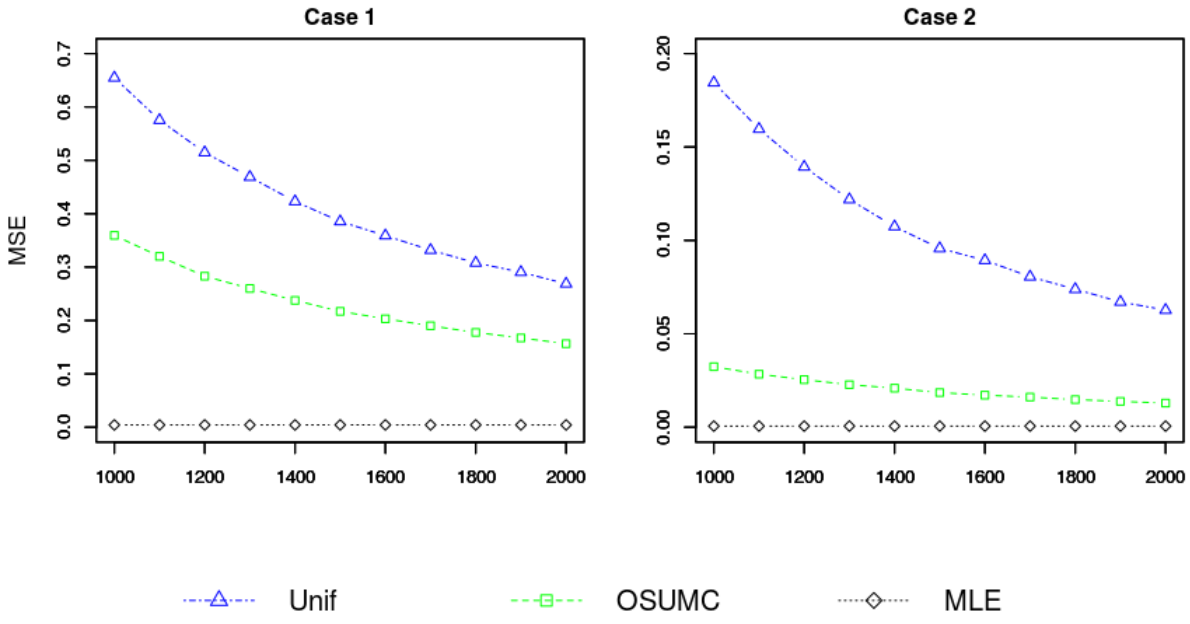



Figure A.6: MSE of the proposed optimal sampling procedure (OSUMC), the uniform sampling (Unif), and the full sample MLE (MLE) for different subsample size $r$ under two scenarios in Poisson regression.

size $n$ and subsample size $r$ in the both considered design generation settings, which again provides empirical support for our theoretical results, especially the discussion of Poisson regression after Theorem 1.

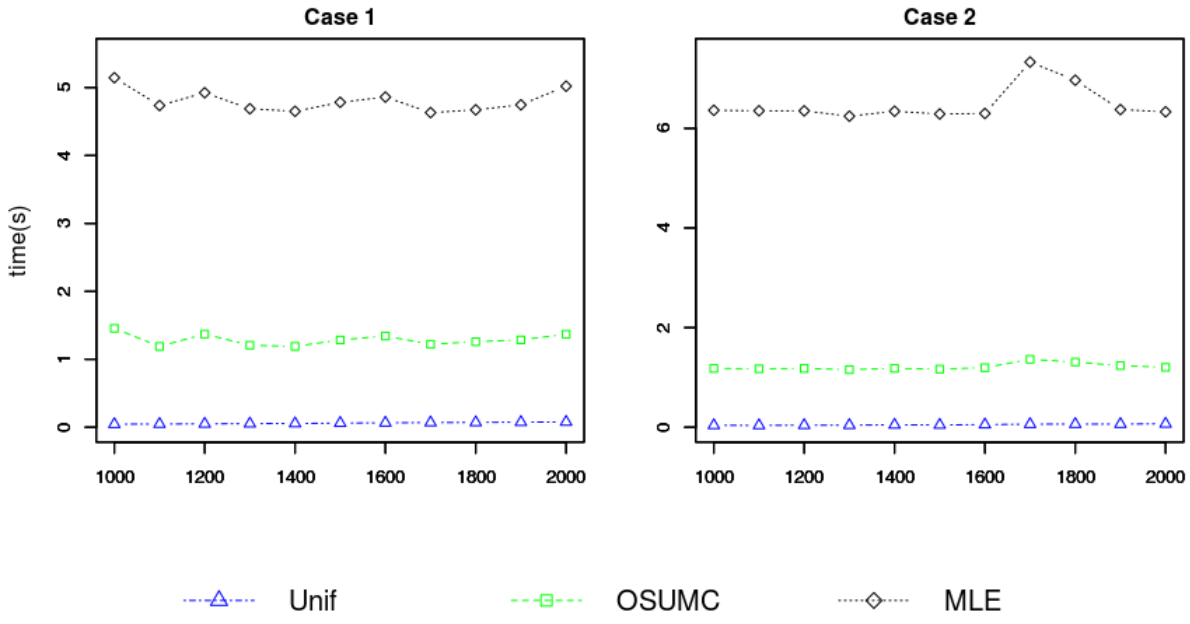
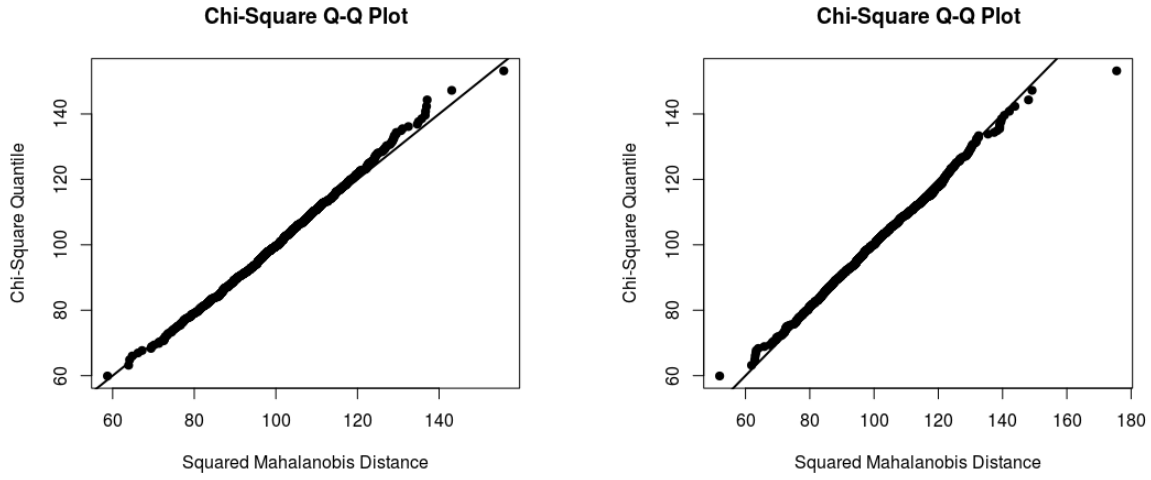



Figure A.7: Computational time plot for different subsample size under different design generation settings for Poisson regression.




(a) Case 1

(b) Case 2

Figure A.8: Chi-square Q-Q plots of $\hat{\beta}_n$ under different design generation settings for Poisson regression with $r = 5000$.

APPENDIX B

**APPENDIX OF CHAPTER 2**

## B.1 Auxiliary results for Section 2.3.3

**Proposition 5** (Limit distribution of maximal deviation)**.** *Suppose that Assumption 1 and Condition (2.6) with $M = L$ hold. Let $\zeta_n := \max_{1\le\ell\le L}\sqrt{nh^3}|\hat{m}(\boldsymbol{x}_\ell) - m(\boldsymbol{x}_\ell)|/\sigma_{\boldsymbol{x}_\ell}$ with $\sigma^2_{\boldsymbol{x}} = \mathbb{E}[\psi_{\boldsymbol{x}}(U,\boldsymbol{X})^2]$. Assume $L = L_n \to \infty$, and define*

$$a_n = (2\log L_n)^{1/2} \quad \textit{and} \quad b_n = (2\log L_n)^{1/2} - \frac{1}{2}(2\log L_n)^{-1/2}(\log\log L_n + \log\pi).$$

*If, in addition, $\tau_{\boldsymbol{x}_1},\dots,\tau_{\boldsymbol{x}_L}$ are all distinct and $\min_{k\neq\ell}|\tau_{\boldsymbol{x}_k} - \tau_{\boldsymbol{x}_\ell}| > 2h$ for sufficiently large $n$, then $a_n(\zeta_n - b_n)$ converges in distribution to the Gumbel distribution, i.e.,*

$$\lim_{n\to\infty}\mathbb{P}(a_n(\zeta_n - b_n)\le t) = e^{-e^{-t}},\ t\in\mathbb{R}.$$

Proposition 5 suggests that we can use the Gumbel approximation to construct simultaneous confidence intervals. The proof shows that if $\min_{k\neq\ell}|\tau_{\boldsymbol{x}_k} - \tau_{\boldsymbol{x}_\ell}| > 2h$, then $\Sigma$ is diagonal so that $\zeta_n$ can be approximated by the maximum in absolute value of $L$ independent $N(0,1)$ random variables, which can be further approximated (after normalization) by the Gumbel distribution by extreme value theory. Compared with the pivotal bootstrap discussed in Section 3.3.2, the Gumbel approximation leads to analytical critical values, so from a computational perspective, using the Gumbel limit seems more attractive. However, the justification of the Gumbel approximation relies on a nontrivial spacing assumption on $\tau_{\boldsymbol{x}_k}$'s (which the pivotal bootstrap does not). More importantly, convergence of normal suprema is known to be extremely slow (Hall, 1991), so simultaneous confidence intervals constructed from the Gumbel approximation may not have desirable coverage accuracy.

The following lemma is useful to establish the coverage guarantee of our confidence intervals (see Example 5 for more discussion).

**Lemma 1.** *Let $Y_n, W_n, Z_n$ be sequences of random variables defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ such that (i) $Y_n$ is measurable relative to a sub-$\sigma$-field $\mathcal{C}_n$ (that may depend on $n$); (ii) $\sup_{t\in\mathbb{R}} |\mathbb{P}(Y_n \leq t) - \mathbb{P}(Z_n \leq t)| \to 0$ and $\sup_{t\in\mathbb{R}} |\mathbb{P}(W_n \leq t \mid \mathcal{C}_n) - \mathbb{P}(Z_n \leq t)| \xrightarrow{P} 0$; (iii) the distribution function of $Z_n$ is continuous for each $n$ ($Z_n$ need not have a limit distribution). Let $\hat{q}_n(\alpha)$ denote the conditional $\alpha$-quantile of $W_n$ given $\mathcal{C}_n$. Then $\mathbb{P}(Y_n \leq \hat{q}_n(\alpha)) \to \alpha$.*

The proofs of the above two auxiliary results can be found in Appendix B.3.3.

## B.2 Technical tools

In this section, we collect technical tools that will be used in the subsequent proofs. For a probability measure $Q$ on a measurable space $(S, \mathcal{S})$ and a class of measurable functions $\mathcal{F}$ on $S$ such that $\mathcal{F} \subset L^2(Q)$, let $N(\mathcal{F}, \|\cdot\|_{Q,2}, \delta)$ denote the $\delta$-covering number for $\mathcal{F}$ with respect to the $L^2(Q)$-seminorm $\|\cdot\|_{Q,2}$. The class $\mathcal{F}$ is said to be pointwise measurable if there exists a countable subclass $\mathcal{G} \subset \mathcal{F}$ such that for every $f \in \mathcal{F}$ there exists a sequence $g_m \in \mathcal{G}$ with $g_m \to f$ pointwise. A function $F : S \to [0, \infty)$ is said to be an envelope for $\mathcal{F}$ if $F(x) \geq \sup_{f\in\mathcal{F}} |f(x)|$ for all $x \in S$. See Section 2.1 in van der Vaart and Wellner (1996) for details. For a vector-valued function $g$ defined over a set $T$, we define $\|g\|_T := \sup_{x\in T} \|g(x)\|$. The $L_{p,1}$ norm for a random variable $X$ is defined as $\|X\|_{p,1} := \int_0^\infty \mathbb{P}(|X| > t)^{1/p} dt$.

**Lemma 2** (Local maximal inequality)**.** *Let $X, X_1, \ldots, X_n$ be i.i.d. random variables taking values in a measurable space $(S, \mathcal{S})$, and let $\mathcal{F}$ be a pointwise measurable class of*

*(measurable) real-valued functions on $S$ with measurable envelope $F$. Suppose that $\mathcal{F}$ is VC type, i.e., there exist constants $A \geq e$ and $V \geq 1$ such that*

$$\sup_Q N(\mathcal{F}, \|\cdot\|_{Q,2}, \epsilon\|F\|_{Q,2}) \leq (A/\epsilon)^V,\ 0 < \forall\epsilon \leq 1,$$

*where $\sup_Q$ is taken over all finitely discrete distributions on $S$. Furthermore, suppose that $0 < \mathbb{E}[F^2(X)] < \infty$, and let $\sigma^2 > 0$ be any positive constant such that $\sup_{f\in\mathcal{F}} \mathbb{E}[f^2(X)] \leq \sigma^2 \leq \mathbb{E}[F^2(X)]$. Define $B = \sqrt{\mathbb{E}[\max_{1\leq i\leq n} F^2(X_i)]}$. Then*

$$\begin{aligned}
&\mathbb{E}\left[\left\|\frac{1}{\sqrt{n}}\sum_{j=1}^n \{f(X_j) - \mathbb{E}[f(X)]\}\right\|_{\mathcal{F}}\right] \\
&\quad \leq C\left[\sqrt{V\sigma^2 \log\left(\frac{A\sqrt{\mathbb{E}[F^2(X)]}}{\sigma}\right)} + \frac{VB}{\sqrt{n}}\log\left(\frac{A\sqrt{\mathbb{E}[F^2(X)]}}{\sigma}\right)\right],
\end{aligned}$$

*where $C > 0$ is a universal constant.*

*Proof.* See Corollary 5.1 in Chernozhukov et al. (2014). □

The following anti-concentration inequality for Gaussian measures (called Nazarov's inequality in Chernozhukov et al. (2017a)), together with the Gaussian comparison inequality, will play crucial roles in proving the validity of the pivotal bootstrap.

**Lemma 3** (Nazarov's inequality)**.** *Let $\boldsymbol{Y} = (Y_1, \ldots, Y_d)^T$ be a centered Gaussian vector in $\mathbb{R}^d$ such that $\mathbb{E}[Y_j^2] \geq \underline{\sigma}^2$ for all $j = 1, \ldots, d$ and some constant $\underline{\sigma} > 0$. Then for every $\boldsymbol{y} \in \mathbb{R}^d$ and $\delta > 0$,*

$$\mathbb{P}(\boldsymbol{Y} \leq \boldsymbol{y} + \delta) - \mathbb{P}(\boldsymbol{Y} \leq \boldsymbol{y}) \leq \frac{\delta}{\underline{\sigma}}(\sqrt{2\log d} + 2).$$

*Proof.* See Lemma A.1 in Chernozhukov et al. (2017a); see also Chernozhukov et al. (2017b). □

**Lemma 4** (Gaussian comparison)**.** *Let $\boldsymbol{Y}$ and $\boldsymbol{W}$ be centered Gaussian random vectors in $\mathbb{R}^d$ with covariance matrices $\Sigma^Y = (\Sigma^Y_{j,k})_{1\leq j,k\leq d}$ and $\Sigma^W = (\Sigma^W_{j,k})_{1\leq j,k\leq d}$, respectively, and let $\Delta = \|\Sigma^Y - \Sigma^W\|_\infty := \max_{1\leq j,k\leq d} |\Sigma^Y_{j,k} - \Sigma^W_{j,k}|$. Suppose that $\min_{1\leq j\leq d} \Sigma^Y_{j,j} \bigvee \min_{1\leq j\leq d} \Sigma^W_{j,j} \geq \underline{\sigma}^2$ for some constant $\underline{\sigma} > 0$. Then*

$$\sup_{b\in\mathbb{R}^d} |\mathbb{P}(\boldsymbol{Y} \leq b) - \mathbb{P}(\boldsymbol{W} \leq b)| \leq C\Delta^{1/3} \log^{2/3} d,$$

*where $C$ is a constant that depends only on $\underline{\sigma}$.*

*Proof.* Implicit in the proof Theorem 4.1 in Chernozhukov et al. (2017a). □

## B.3 Proofs for Section 2.3

### B.3.1 Uniform Convergence Rates

We first establish uniform convergence rates of $\hat{Q}^{(r)}_{\boldsymbol{x}}(\tau)$. The following Bahadur representation of the linear quantile regression estimator $\hat{\beta}(\tau)$ will be used in the subsequent proofs.

**Lemma 5** (Bahadur representation of $\hat{\beta}(\tau)$)**.** *Under Assumption 1, we have*

$$\hat{\beta}(\tau) - \beta(\tau) = J(\tau)^{-1} \left[\frac{1}{n}\sum_{i=1}^n \{\tau - I(U_i \leq \tau)\}\boldsymbol{X}_i\right] + o_P(n^{-3/4}\log n),$$

*uniformly in $\tau \in [\epsilon/2, 1-\epsilon/2]$, where $U_1, \ldots, U_n \sim U(0,1)$ i.i.d. that are independent of $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$. In addition, we have*

$$\sup_{\tau\in[\epsilon/2,1-\epsilon/2]} \left\| \frac{1}{n}\sum_{i=1}^n \{\tau - I(U_i \leq \tau)\}\boldsymbol{X}_i \right\| = O_P(n^{-1/2}).$$

*Proof.* See Lemma 3 in Ota et al. (2019). See also Ruppert and Carroll (1980); Gutenbrunner and Jurecková (1992); He and Shao (1996). □

**Remark 12.** Inspection of the proof shows that $U_i = F(Y_i \mid \boldsymbol{X}_i)$ where $F(y \mid \boldsymbol{X})$ is the conditional distribution function of $Y$ given $\boldsymbol{X}$.

We first prove the following technical lemma.

**Lemma 6.** *If Assumption 1 holds, then for $r = 1, 2, 3$, we have*

$$\sup_{\substack{\boldsymbol{x}\in\mathcal{X}_0 \\ \tau\in[\epsilon,1-\epsilon]}} \left| \frac{1}{nh^r} \sum_{i=1}^{n} \boldsymbol{x}^T J(\tau)^{-1} \boldsymbol{X}_i \left\{ K^{(r-1)}\left(\frac{\tau - U_i}{h}\right) - hI(r=1) \right\} \right| = O_P\left(n^{-1/2}h^{-r+1/2}\sqrt{\log n}\right).$$

*Proof.* Since $K$ is supported in $[-1, 1]$, for sufficiently large $n$,

$$\mathbb{E}\left[K^{(r-1)}\left(\frac{\tau - U}{h}\right)\right] = h\int_{(1-\tau)/h}^{\tau/h} K^{(r-1)}(t)dt = h\int_{\mathbb{R}} K^{(r-1)}(t)dt = hI(r=1).$$

Consider the function class $\mathcal{F}_h := \{(u, \boldsymbol{x}') \mapsto K^{(r-1)}((\tau - u)/h)\boldsymbol{x}^T J(\tau)^{-1}\boldsymbol{x}' : \boldsymbol{x} \in \mathcal{X}_0, \tau \in [\epsilon, 1-\epsilon]\}$ (which depends on $n$ since $h = h_n$ does). It suffices to show that

$$\mathbb{E}[\|\mathbb{G}_n\|_{\mathcal{F}_h}] = O(\sqrt{h\log n}) \quad \text{with} \quad \mathbb{G}_n f = n^{-1/2}\sum_{i=1}^{n}\{f(U_i, \boldsymbol{X}_i) - \mathbb{E}[f(U, \boldsymbol{X})]\}.$$

To this end, we will apply Lemma 2. The function class $\mathcal{F}_h$ is a subset of the pointwise product of the following two function classes (that are independent of $n$): $\mathcal{F}' = \{(u, \boldsymbol{x}') \mapsto \boldsymbol{x}^T J(\tau)^{-1}\boldsymbol{x}' : \boldsymbol{x} \in \mathcal{X}_0,\ \tau \in [\epsilon, 1-\epsilon]\}$ and $\mathcal{F}'' = \{(u, \boldsymbol{x}') \mapsto K^{(r-1)}(au + b) : a, b \in \mathbb{R}\}$. The former function class $\mathcal{F}'$ has envelope $F_1(u, \boldsymbol{x}') = C\|\boldsymbol{x}'\|$ and the latter function class $\mathcal{F}''$ has envelope $F_2(u, \boldsymbol{x}') = C'$ where $C, C'$ are some constants independent of $n$. The function class $\mathcal{F}'$ is a subset of a vector space of dimension $d$, so that it is a VC subgraph class with VC index at most $d + 2$ (cf. Lemma 2.6.15 in van der Vaart and Wellner (1996)). Next, since $K^{(r-1)}$ is of bounded variation (i.e., it can be written as the difference of two bounded nondecreasing functions) and the function class $\{u \mapsto au + b : a, b \in \mathbb{R}\}$ is a VC subgraph class (as it is a vector space of dimension 2), the function class $\mathcal{F}''$ is

VC type in view of Lemma 2.6.18 in van der Vaart and Wellner (1996). Conclude that, for $F(u, \boldsymbol{x}') = CC'\|\boldsymbol{x}'\|$, there exist positive constants $A, V$ independent of $n$ such that

$$\sup_Q N(\mathcal{F}_h, \|\cdot\|_{Q,2}, \eta\|F\|_{Q,2}) \leq (A/\eta)^V,\ 0 < \forall\eta \leq 1,$$

where $\sup_Q$ is taken over all finitely discrete distributions on $(0,1) \times \mathbb{R}^d$.

It is not difficult to verify that, by independence between $U$ and $\boldsymbol{X}$,

$$\sup_{\substack{\boldsymbol{x}\in\mathcal{X}_0 \\ \tau\in[\epsilon,1-\epsilon]}} \mathbb{E}[\{K^{(r-1)}((\tau-U)/h)\boldsymbol{x}^T J(\tau)^{-1}\boldsymbol{X}\}^2] \leq O(1)\int_0^1 K^{(r-1)}((\tau-u)/h)^2 du = O(h).$$

In addition, $\mathbb{E}[\max_{1\leq i\leq n} F^2(U_i, \boldsymbol{X}_i)] \leq O(1)\mathbb{E}[\max_{1\leq i\leq n}\|\boldsymbol{X}_i\|^2] = O(n^{1/2})$ (as $\mathbb{E}[\|\boldsymbol{X}\|^4] < \infty$). Conclude from Lemma 2 that

$$\mathbb{E}[\|\mathbb{G}_n\|_{\mathcal{F}_h}] = O(\sqrt{h\log n} + n^{-1/4}\log n) = O(\sqrt{h\log n}).$$

This completes the proof. □

The following lemma derives uniform convergence rates of $\hat{Q}_{\boldsymbol{x}}^{(r)}(\tau)$.

**Lemma 7** (Uniform convergence rates $\hat{Q}_{\boldsymbol{x}}^{(r)}(\tau)$)**.** *Under Assumption 1, we have*

$$\sup_{\substack{\boldsymbol{x}\in\mathcal{X}_0 \\ \tau\in[\epsilon,1-\epsilon]}} |\hat{Q}_{\boldsymbol{x}}^{(r)}(\tau) - Q_{\boldsymbol{x}}^{(r)}(\tau)| = \begin{cases} O_P\left(n^{-1/2} + h^2\right) & \text{if } r = 0 \\ O_P\left(n^{-1/2}h^{-r+1/2}\sqrt{\log n} + h^2\right) & \text{if } r = 1 \text{ or } 2 \\ O_P\left(n^{-1/2}h^{-5/2}\sqrt{\log n} + h\right) & \text{if } r = 3 \end{cases}$$

*Proof.* Consider first the case where $r = 0$. By definition,

$$\begin{aligned}\sup_{\substack{\boldsymbol{x}\in\mathcal{X}_0\\ \tau\in[\epsilon,1-\epsilon]}} |\hat{Q}_{\boldsymbol{x}}(\tau) - Q_{\boldsymbol{x}}(\tau)| &= \sup_{\substack{\boldsymbol{x}\in\mathcal{X}_0\\ \tau\in[\epsilon,1-\epsilon]}} \left|\int \check{Q}_{\boldsymbol{x}}(t) K_h(\tau - t)dt - Q_{\boldsymbol{x}}(\tau)\right| \\ &\leq \sup_{\substack{\boldsymbol{x}\in\mathcal{X}_0\\ \tau\in[\epsilon,1-\epsilon]}} \left|\int [\check{Q}_{\boldsymbol{x}}(t) - Q_{\boldsymbol{x}}(t)] K_h(\tau - t)dt\right| \\ &\quad + \sup_{\substack{\boldsymbol{x}\in\mathcal{X}_0\\ \tau\in[\epsilon,1-\epsilon]}} \left|\int Q_{\boldsymbol{x}}(t) K_h(\tau - t)dt - Q_{\boldsymbol{x}}(\tau)\right| \\ &=: I + II.\end{aligned}$$

We have $I = O_P(n^{-1/2})$ by Lemma 5 and $II = O(h^2)$ by Taylor expansion.

Next, consider $1 \leq r \leq 3$. We note that

$$\begin{aligned}\sup_{\substack{\boldsymbol{x}\in\mathcal{X}_0\\ \tau\in[\epsilon,1-\epsilon]}} |\hat{Q}^{(r)}_{\boldsymbol{x}}(\tau) - Q^{(r)}_{\boldsymbol{x}}(\tau)| &\leq \sup_{\substack{\boldsymbol{x}\in\mathcal{X}_0\\ \tau\in[\epsilon,1-\epsilon]}} \left|\int [\check{Q}_{\boldsymbol{x}}(t) - Q_{\boldsymbol{x}}(t)] K^{(r)}_h(\tau - t)dt\right| \\ &\quad + \sup_{\substack{\boldsymbol{x}\in\mathcal{X}_0\\ \tau\in[\epsilon,1-\epsilon]}} \left|\int Q_{\boldsymbol{x}}(t) K^{(r)}_h(\tau - t)dt - Q^{(r)}_{\boldsymbol{x}}(\tau)\right| \\ &=: III + IV.\end{aligned}$$

We have $IV = O(h^2)$ for $r = 1, 2$ and $= O(h)$ for $r = 3$ by Taylor expansion (recall that $Q_{\boldsymbol{x}}(\tau)$ is four-times continuously differentiable). Observe that, by Lemma 5 and change of variables,

$$\begin{aligned}III \leq \sup_{\substack{\boldsymbol{x}\in\mathcal{X}_0\\ \tau\in[\epsilon,1-\epsilon]}} &\left|\frac{1}{nh^r}\sum_{i=1}^{n}\int \boldsymbol{x}^T J(\tau - th)^{-1}\boldsymbol{X}_i \left\{\tau - th - I\left(U_i \leq \tau - th\right)\right\} K^{(r)}(t)dt\right| \\ &+ \underbrace{o_P(n^{-3/4}h^{-r}\log n)}_{=o_P(n^{-1/2}h^{-r+1/2}\sqrt{\log n})}.\end{aligned}$$

Replacing $J(\tau - th)$ by $J(\tau)$ in the first term on the right hand side results in an error of order $O_P(n^{-1/2}h^{-r+1})$; this can be verified by a similar argument to the

proof of the preceding lemma. Thus, it remains to bound

$$\sup_{\substack{\boldsymbol{x}\in\mathcal{X}_0\\ \tau\in[\epsilon,1-\epsilon]}}\left|\frac{1}{nh^r}\sum_{i=1}^{n}\int \boldsymbol{x}^T J(\tau)^{-1}\boldsymbol{X}_i\left\{\tau-th-I\left(U_i\le\tau-th\right)\right\}K^{(r)}\left(t\right)dt\right|$$
$$=\sup_{\substack{\boldsymbol{x}\in\mathcal{X}_0\\ \tau\in[\epsilon,1-\epsilon]}}\left|\frac{1}{nh^r}\sum_{i=1}^{n}\boldsymbol{x}^T J(\tau)^{-1}\boldsymbol{X}_i\left\{K^{(r-1)}\left(\frac{\tau-U_i}{h}\right)+h\int tK^{(r)}\left(t\right)dt\right\}\right|,$$

where we have used the fact that $K^{(r)}$ integrates to $0$. Here, by integration by parts,

$$\int tK^{(r)}(t)dt=-\int K^{(r-1)}(t)dt=-I(r=1).$$

Thus, from Lemma 6, we have $III=O(n^{-1/2}h^{-r+1/2}\sqrt{\log n})$. This completes the proof. □

**Remark 13** (Bias of $\hat{Q}_{\boldsymbol{x}}(\tau)$ at $\tau=\tau_{\boldsymbol{x}}$)**.** The bias of $\hat{Q}_{\boldsymbol{x}}(\tau)$ can be improved to $O(h^4)$ at $\tau=\tau_{\boldsymbol{x}}$ by $Q''_{\boldsymbol{x}}(\tau_{\boldsymbol{x}})=0$ and symmetry of $K$.

**Remark 14** (Expansion of $\hat{Q}''_{\boldsymbol{x}}(\tau)$)**.** Inspection of the proof shows that

$$\begin{aligned}\hat{Q}''_{\boldsymbol{x}}(\tau)-Q''_{\boldsymbol{x}}(\tau)-\frac{Q^{(4)}_{\boldsymbol{x}}(\tau)}{2}\kappa h^2+o(h^2)&=\frac{1}{nh^2}\sum_{i=1}^{n}K'\left(\frac{\tau-U_i}{h}\right)\boldsymbol{x}^T J(\tau)^{-1}\boldsymbol{X}_i\\ &+O_P(n^{-1/2}h^{-1})+\underbrace{o_P(n^{-3/4}h^{-2}\log n)}_{o_P(n^{-1/2}h^{-1})}\end{aligned}\tag{B.1}$$

uniformly in $(\tau,\boldsymbol{x})\in[\epsilon,1-\epsilon]\times\mathcal{X}_0$. Recall that $\kappa=\int t^2K(t)dt$.

## B.3.2 Proofs for Section 2.3.2

We first prove the uniform consistency of $\hat{\tau}_{\boldsymbol{x}}$.

**Lemma 8** (Uniform consistency of $\hat{\tau}_{\boldsymbol{x}}$)**.** *Under Assumption 1, we have* $\sup_{\boldsymbol{x}\in\mathcal{X}_0}|\hat{\tau}_{\boldsymbol{x}}-\tau_{\boldsymbol{x}}|\xrightarrow{P}0$.

*Proof.* We divide the proof into two steps.

**Step 1**. We will verify that for any $\delta > 0$,

$$\eta_\delta := \inf_{\boldsymbol{x} \in \mathcal{X}_0} \inf_{\substack{\tau \in [\epsilon, 1-\epsilon] \\ |\tau - \tau_{\boldsymbol{x}}| \geq \delta}} \{s_{\boldsymbol{x}}(\tau) - s_{\boldsymbol{x}}(\tau_{\boldsymbol{x}})\} > 0.$$

This follows from the following two claims: (i) $s_{\boldsymbol{x}}(\tau) - s_{\boldsymbol{x}}(\tau_{\boldsymbol{x}})$ is jointly continuous in $(\tau, \boldsymbol{x})$, (ii) $S_\delta := \{(\tau, \boldsymbol{x}) : \boldsymbol{x} \in \mathcal{X}_0, \tau \in [\epsilon, 1-\epsilon], |\tau - \tau_{\boldsymbol{x}}| \geq \delta\}$ is compact in $(0,1) \times \mathbb{R}^d$ and the observation that $\tau_{\boldsymbol{x}}$ is the unique minimizer of $s_{\boldsymbol{x}}(\tau)$, i.e., $\tau_{\boldsymbol{x}} = \arg\min_{\tau \in [\epsilon, 1-\epsilon]} s_{\boldsymbol{x}}(\tau)$. Since $s_{\boldsymbol{x}}(\tau) = \partial Q_{\boldsymbol{x}}(\tau)/\partial \tau$ is continuous in $\tau$ for any fixed $\boldsymbol{x}$ under Assumption 1 and also linear (thus convex) in $\boldsymbol{x}$ by the linear quantile assumption, Theorem 10.7 in Rockafellar (1970) implies that $s_{\boldsymbol{x}}(\tau)$ is jointly continuous in $(\tau, \boldsymbol{x})$. Now, by Berge's maximum theorem (cf. Theorem 17.31 in Aliprantis and Border (2006): see also their Lemma 17.6), we see that $\tau_{\boldsymbol{x}}$ is continuous in $\boldsymbol{x}$. The preceding discussion also implies that $s_{\boldsymbol{x}}(\tau) - s_{\boldsymbol{x}}(\tau_{\boldsymbol{x}})$ is jointly continuous in $(\tau, \boldsymbol{x})$. Combining the continuity of $\tau_{\boldsymbol{x}}$ and the definition of $S_\delta$, we can verify $S_\delta$ is closed and bounded and therefore compact. Thus, we have verified claims (i) and (ii) and the conclusion of this step follows.

**Step 2**. We will prove the uniform consistency of $\hat{\tau}_{\boldsymbol{x}}$. Consider the event $\mathcal{A}_\delta := \{\sup_{\boldsymbol{x} \in \mathcal{X}_0} \{s_{\boldsymbol{x}}(\hat{\tau}_{\boldsymbol{x}}) - s_{\boldsymbol{x}}(\tau_{\boldsymbol{x}})\} \geq \eta_\delta\}$. Observe that

$$\begin{aligned}
&\sup_{\boldsymbol{x} \in \mathcal{X}_0} \{s_{\boldsymbol{x}}(\hat{\tau}_{\boldsymbol{x}}) - s_{\boldsymbol{x}}(\tau_{\boldsymbol{x}})\} \\
&\qquad \leq \sup_{\boldsymbol{x} \in \mathcal{X}_0} \{s_{\boldsymbol{x}}(\hat{\tau}_{\boldsymbol{x}}) - \hat{s}_{\boldsymbol{x}}(\hat{\tau}_{\boldsymbol{x}})\} + \sup_{\boldsymbol{x} \in \mathcal{X}_0} \{\hat{s}_{\boldsymbol{x}}(\hat{\tau}_{\boldsymbol{x}}) - \hat{s}_{\boldsymbol{x}}(\tau_{\boldsymbol{x}})\} + \sup_{\boldsymbol{x} \in \mathcal{X}_0} \{\hat{s}_{\boldsymbol{x}}(\tau_{\boldsymbol{x}}) - s_{\boldsymbol{x}}(\tau_{\boldsymbol{x}})\}.
\end{aligned}$$

The first and third terms on the right hand side are $o_P(1)$ by Lemma 7, while the second term is nonpositive by the definition of $\hat{\tau}_{\boldsymbol{x}}$. This implies that $\mathbb{P}(\mathcal{A}_\delta) \leq \mathbb{P}(\eta_\delta \leq o_P(1)) = o(1)$. The uniform consistency of $\hat{\tau}_{\boldsymbol{x}}$ follows from the fact that the event $\{\sup_{\boldsymbol{x} \in \mathcal{X}_0} |\hat{\tau}_{\boldsymbol{x}} - \tau_{\boldsymbol{x}}| \geq \delta\}$ is included in $\mathcal{A}_\delta$. □

The uniform consistency guarantees that the first order condition for $\hat{\tau}_{\boldsymbol{x}}$ holds for all $\boldsymbol{x} \in \mathcal{X}_0$ with probability approaching one, i.e.,

$$\mathbb{P}\left(\hat{s}'_{\boldsymbol{x}}(\hat{\tau}_{\boldsymbol{x}}) = 0, \forall \boldsymbol{x} \in \mathcal{X}_0\right) \to 1. \tag{B.2}$$

Recall that $\hat{s}'_{\boldsymbol{x}}(\tau) = \hat{Q}''_{\boldsymbol{x}}(\tau)$. Now, we derive an asymptotic linear representation for $\hat{\tau}_{\boldsymbol{x}}$.

**Lemma 9** (Asymptotic linear representation of $\hat{\tau}_{\boldsymbol{x}}$)**.** *Under Assumption 1, the following expansion holds uniformly in $\boldsymbol{x} \in \mathcal{X}_0$:*

$$\begin{aligned}
&\hat{\tau}_{\boldsymbol{x}} - \tau_{\boldsymbol{x}} + \frac{s^{(3)}_{\boldsymbol{x}}(\tau_{\boldsymbol{x}})}{2s''_{\boldsymbol{x}}(\tau_{\boldsymbol{x}})}\kappa h^2 + o_P(h^2) \\
&\quad = -\frac{1}{nh^2 s''_{\boldsymbol{x}}(\tau_{\boldsymbol{x}})} \sum_{i=1}^{n} K'\left(\frac{\tau_{\boldsymbol{x}} - U_i}{h}\right) \boldsymbol{x}^T J(\tau_{\boldsymbol{x}})^{-1} \boldsymbol{X}_i + O_P(n^{-1/2}h^{-1} + n^{-1}h^{-4}\log n).
\end{aligned}$$

*In addition, the first term on the right hand side is $O_P(n^{-1/2}h^{3/2}\sqrt{\log n})$ uniformly in $\boldsymbol{x} \in \mathcal{X}_0$.*

*Proof.* From the first order condition (B.2) coupled with the Taylor expansion, we have

$$0 = \hat{Q}''_{\boldsymbol{x}}(\hat{\tau}_{\boldsymbol{x}}) = \hat{Q}''_{\boldsymbol{x}}(\tau_{\boldsymbol{x}}) + \hat{Q}^{(3)}_{\boldsymbol{x}}(\check{\tau}_{\boldsymbol{x}})(\hat{\tau}_{\boldsymbol{x}} - \tau_{\boldsymbol{x}}).$$

where $\check{\tau}_{\boldsymbol{x}}$ lies between $\hat{\tau}_{\boldsymbol{x}}$ and $\tau_{\boldsymbol{x}}$. This yields that

$$\hat{\tau}_{\boldsymbol{x}} - \tau_{\boldsymbol{x}} = -\hat{Q}^{(3)}_{\boldsymbol{x}}(\check{\tau}_{\boldsymbol{x}})^{-1} \cdot \hat{Q}''_{\boldsymbol{x}}(\tau_{\boldsymbol{x}}).$$

The rest of the proof is divided into two steps.

**Step 1**. We will show that $\sup_{\boldsymbol{x}\in\mathcal{X}_0} |\hat{\tau}_{\boldsymbol{x}} - \tau_{\boldsymbol{x}}| = O_P(n^{-1/2}h^{-3/2}\sqrt{\log n} + h^2)$. Observe that $\hat{Q}^{(3)}_{\boldsymbol{x}}(\check{\tau}_{\boldsymbol{x}}) = Q^{(3)}_{\boldsymbol{x}}(\check{\tau}_{\boldsymbol{x}}) + O_P(n^{-1/2}h^{-5/2}\sqrt{\log n} + h) = Q^{(3)}_{\boldsymbol{x}}(\check{\tau}_{\boldsymbol{x}}) + o_P(1)$ uniformly in $\boldsymbol{x} \in \mathcal{X}_0$ by Lemma 6, $Q^{(3)}_{\boldsymbol{x}}(\check{\tau}_{\boldsymbol{x}}) = Q^{(3)}_{\boldsymbol{x}}(\tau_{\boldsymbol{x}}) + o_P(1)$ uniformly in $\boldsymbol{x} \in \mathcal{X}_0$ by the uniform consistency of $\hat{\tau}_{\boldsymbol{x}}$, and the map $\boldsymbol{x} \mapsto Q^{(3)}_{\boldsymbol{x}}(\tau_{\boldsymbol{x}})$ is bounded

away from zero on $\mathcal{X}_0$. Thus, we have

$$\sup_{\boldsymbol{x}\in\mathcal{X}_0} |\hat{\tau}_{\boldsymbol{x}} - \tau_{\boldsymbol{x}}| = O_P\left(\sup_{\boldsymbol{x}\in\mathcal{X}_0} |\hat{Q}''_{\boldsymbol{x}}(\tau_{\boldsymbol{x}})|\right).$$

However, since $Q''_{\boldsymbol{x}}(\tau_{\boldsymbol{x}}) = 0$, the right hand side on the above equation is $O_P(n^{-1/2}h^{-3/2}\sqrt{\log n} + h^2)$ by Lemma 6.

**Step 2**. We wish to derive the conclusion of the lemma. From the preceding discussion, we see that $\hat{Q}^{(3)}_{\boldsymbol{x}}(\check{\tau}_{\boldsymbol{x}}) = Q^{(3)}_{\boldsymbol{x}}(\tau_{\boldsymbol{x}}) + O_P(n^{-1/2}h^{-5/2}\sqrt{\log n} + h)$ uniformly in $\boldsymbol{x} \in \mathcal{X}_0$, so that

$$\hat{\tau}_{\boldsymbol{x}} - \tau_{\boldsymbol{x}} = -Q^{(3)}_{\boldsymbol{x}}(\tau_{\boldsymbol{x}})^{-1}\hat{Q}''_{\boldsymbol{x}}(\tau_{\boldsymbol{x}}) + O_P(n^{-1}h^{-4}\log n + n^{-1/2}h^{-1/2}\sqrt{\log n} + h^3)$$

uniformly in $x \in \mathcal{X}_0$. The conclusion of the lemma follows from combining the expansion (B.1). □

We are now in position to prove Proposition 1.

*Proof of Proposition 1.* We note that, uniformly in $\boldsymbol{x} \in \mathcal{X}_0$,

$$\begin{aligned}
&\hat{m}(\boldsymbol{x}) - m(\boldsymbol{x}) \\
&\quad = \{\hat{Q}_{\boldsymbol{x}}(\hat{\tau}_{\boldsymbol{x}}) - Q_{\boldsymbol{x}}(\hat{\tau}_{\boldsymbol{x}})\} + \{Q_{\boldsymbol{x}}(\hat{\tau}_{\boldsymbol{x}}) - Q_{\boldsymbol{x}}(\tau_{\boldsymbol{x}})\} \\
&\quad = Q_{\boldsymbol{x}}(\hat{\tau}_{\boldsymbol{x}}) - Q_{\boldsymbol{x}}(\tau_{\boldsymbol{x}}) + O_P(n^{-1/2}) + o_P(h^2) \quad \text{(by Lemma 6 and Remark 13)} \\
&\quad = Q'_{\boldsymbol{x}}(\tau_{\boldsymbol{x}})(\hat{\tau}_{\boldsymbol{x}} - \tau_{\boldsymbol{x}}) + O_P\left(\sup_{\boldsymbol{x}'\in\mathcal{X}_0} |\hat{\tau}_{\boldsymbol{x}'} - \tau_{\boldsymbol{x}'}|^3\right) + O_P(n^{-1/2}) + o_P(h^2) \quad \text{(by } Q''_{\boldsymbol{x}}(\tau_{\boldsymbol{x}}) = 0\text{)} \\
&\quad = \underbrace{\frac{1}{nh^{3/2}}\sum_{i=1}^{n}\psi_{\boldsymbol{x}}(U_i, \boldsymbol{X}_i)}_{=O_P(n^{-1/2}h^{-3/2}\sqrt{\log n})} - \frac{s_{\boldsymbol{x}}(\tau_{\boldsymbol{x}})s^{(3)}_{\boldsymbol{x}}(\tau_{\boldsymbol{x}})}{2s''(\tau_{\boldsymbol{x}})}\kappa h^2 \\
&\qquad + O_P(n^{-1/2}h^{-1} + n^{-1}h^{-4}\log n) + o_P(h^2). \quad \text{(by Lemma 9)}
\end{aligned}$$

This completes the proof. □

*Proof of Corollary 1.* Proposition 1 implies that, for any fixed $\boldsymbol{x} \in \mathcal{X}_0$,

$$\hat{m}(\boldsymbol{x}) - m(\boldsymbol{x}) = \frac{1}{nh^{3/2}} \sum_{i=1}^{n} \psi_{\boldsymbol{x}}(U_i, \boldsymbol{X}_i) + o_P(n^{-1/2}h^{-3/2}).$$

Thus, it suffices to show that $n^{-1/2} \sum_{i=1}^{n} \psi_{\boldsymbol{x}}(U_i, \boldsymbol{X}_i) \xrightarrow{d} N(0, V_{\boldsymbol{x}})$. Recall that $\psi_{\boldsymbol{x}}(U_i, \boldsymbol{X}_i)$ has mean zero. The above result follows from verifying the Lyapunov condition, together with the fact that $\mathbb{E}[\psi_{\boldsymbol{x}}(U, \boldsymbol{X})^2] = V_{\boldsymbol{x}}$. We omit the details for brevity. □

### B.3.3 Proofs for Section 2.3.3

#### Proof of Theorem 1

We divide the proof into two steps.

**Step 1**. We will show that

$$\sup_{b \in \mathbb{R}^M} \left| \mathbb{P}\left( n^{-1/2} \textstyle\sum_{i=1}^{n} A\Psi_i \leq b \right) - \mathbb{P}\left( AG \leq b \right) \right| \to 0. \tag{B.3}$$

To this end, we verify Conditions (M.1), (M.2), and (E.2) in Proposition 2.1 of Chernozhukov et al. (2017a).

**Condition (M.1)**: For $k = 1, \ldots, M$, by definition and Assumption 2,

$$\mathbb{E}[(A_k^T \Psi_i)^2] = A_k^T \mathbb{E}[\Psi_i \Psi_i^T] A_k = D_k^T \Sigma D_k / \Gamma_k^2,$$

which is bounded away from zero uniformly over $1 \leq k \leq M$.

**Condition (M.2)**: For $k = 1, \ldots, M$,

$$\begin{aligned}
&\mathbb{E}\left[ \left| A_k^T \Psi_i \right|^3 \right] \leq \max_{\ell \in S_k} |A_{k,\ell}|^3 \mathbb{E}\left[ \| (\psi_{\boldsymbol{x}_\ell}(U, \boldsymbol{X}))_{\ell \in S_k} \|_1^3 \right] \\
&\quad \leq \max_{\ell \in S_k} |A_{k,\ell}|^3 \cdot |S_k|^2 \sum_{\ell \in S_k} \mathbb{E}[|\psi_{\boldsymbol{x}_\ell}(U, \boldsymbol{X})|^3] \leq \max_{\ell \in S_k} |A_{k,\ell}|^3 \cdot |S_k|^3 \max_{1 \leq \ell \leq L} \mathbb{E}[|\psi_{\boldsymbol{x}_\ell}(U, \boldsymbol{X})|^3]
\end{aligned}$$

Under our assumption, $\max_{1\leq k\leq M;\ell\in S_k}|A_{k,\ell}| = O(1)$ and $\max_{1\leq k\leq M}|S_k| = O(1)$. In addition,

$$\begin{aligned}&\max_{1\leq\ell\leq L}\mathbb{E}[|\psi_{\boldsymbol{x}_\ell}(U,\boldsymbol{X})|^3]\\&\quad\leq O(h^{-3/2})\mathbb{E}[\|\boldsymbol{X}\|^3]\max_{1\leq\ell\leq L}\int_0^1\left|K'\left(\frac{\tau_{\boldsymbol{x}_\ell}-u}{h}\right)\right|^3du = O(h^{-1/2}).\end{aligned}$$

Likewise, we have $\max_{1\leq k\leq M}\mathbb{E}[|A_k^T\Psi_i|^4] = O(h^{-1})$.

**Condition (E.2)**: Similarly to the previous case (but bounding $h^{-1/2}K'((\tau_{\boldsymbol{x}_\ell}-U)/h)$ by $h^{-1/2}\|K'\|_\infty$), we can show that

$$\mathbb{E}\left[\max_{1\leq k\leq M}\left|A_k^T\Psi_i\right|^q\right]\leq O(h^{-q/2})\mathbb{E}[\|\boldsymbol{X}\|^q] = O(h^{-q/2}).$$

Thus, we can apply Proposition 2.1 in Chernozhukov et al. (2017a), and the conclusion of this step follows as soon as

$$\frac{\log^7(Mn)}{nh}\bigvee\frac{\log^3(Mn)}{n^{1-2/q}h}\to 0,$$

but this is satisfied under our assumption.

**Step 2**. Define $\delta_n = h^{1/2}+n^{-1/2}h^{-5/2}\log n+n^{1/2}h^{7/2}$ and $R_n = (\hat{m}(\boldsymbol{x}_\ell)-m(\boldsymbol{x}_\ell))_{\ell=1}^L-(nh^{3/2})^{-1}\sum_{i=1}^n\Psi_i$. By Proposition 1, we know that $\sqrt{nh^3}\|R_n\|_\infty = O_P(\delta_n)$, so that

$$\begin{aligned}\sqrt{nh^3}\|AR_n\|_\infty&\leq\max_{1\leq k\leq M}\sum_{\ell\in S_k}|A_{k,\ell}||\sqrt{nh^3}R_{n,\ell}|\leq\max_{1\leq k\leq M;1\leq\ell\leq L}|A_{k,\ell}|\max_{1\leq\ell\leq L}\sum_{\ell\in S_k}|\sqrt{nh^3}R_{n,\ell}|\\&\leq\max_{1\leq k\leq M;1\leq\ell\leq L}|A_{k,\ell}|\max_{1\leq k\leq M}|S_k|\|\sqrt{nh^3}R_n\|_\infty = O_P(\delta_n).\end{aligned}$$

Thus, for any $B_n\to\infty$, we have $\mathbb{P}(\sqrt{nh^3}\|AR_n\|_\infty > B_n\delta_n) = o(1)$. Now, for any

$b \in \mathbb{R}^M$,

$$\mathbb{P}\left(A\sqrt{nh^3}(\hat{m}(\boldsymbol{x}_\ell) - m(\boldsymbol{x}_\ell))_{\ell=1}^L \le b\right)$$
$$\le \mathbb{P}\left(n^{-1/2}\textstyle\sum_{i=1}^n A\Psi_i \le b + B_n\delta_n\right) + o(1)$$
$$\le \mathbb{P}(AG \le b + B_n\delta_n) + o(1) \quad \text{(by Step 1)}$$
$$\le \mathbb{P}(AG \le b) + O(B_n\delta_n\sqrt{\log M}) + o(1), \quad \text{(by Nazarov's inequality (Lemma 3))}$$

where the $o$ and $O$ terms are independent of $b$. Likewise, we have

$$\mathbb{P}\left(A\sqrt{nh^3}(\hat{m}(\boldsymbol{x}_\ell) - m(\boldsymbol{x}_\ell))_{\ell=1}^L \le b\right) \ge \mathbb{P}(AG \le b) - O(B_n\delta_n\sqrt{\log M}) - o(1).$$

Since $B_n\delta_n\sqrt{\log M} \to 0$ for sufficiently slow $B_n \to \infty$ under our assumption, we obtain the conclusion of the theorem. □

**Proofs of Theorem 2 and Proposition 2**

We start with proving some technical lemmas. We use $\|\cdot\|_{\mathrm{op}}$ to denote the operator norm of a matrix.

**Lemma 10.** *Under Assumption 1, we have*

$$\sup_{\boldsymbol{x}\in\mathcal{X}_0} \|\hat{J}(\hat{\tau}_{\boldsymbol{x}}) - J(\tau_{\boldsymbol{x}})\|_{\mathrm{op}} = O_P(n^{-1/2}h^{-3/2}\sqrt{\log n} + h^2).$$

*Proof.* It suffices to show that $\sup_{\boldsymbol{x}\in\mathcal{X}_0} |\hat{J}_{j,k}(\hat{\tau}_{\boldsymbol{x}}) - J_{j,k}(\tau_{\boldsymbol{x}})| = O_P(n^{-1/2}h^{-3/2}\sqrt{\log n} + h^2)$ for any $1 \le j, k \le d$ (as the dimension $d$ is fixed). Observe that

$$\begin{aligned}
&\sup_{\boldsymbol{x}\in\mathcal{X}_0} |\hat{J}_{j,k}(\hat{\tau}_{\boldsymbol{x}}) - J_{j,k}(\tau_{\boldsymbol{x}})| \\
&\le \sup_{\boldsymbol{x}\in\mathcal{X}_0} \left| \frac{1}{n}\sum_{i=1}^n K_h(Y_i - \boldsymbol{X}_i^T\hat{\beta}(\hat{\tau}_{\boldsymbol{x}}))X_{ij}X_{ik} - \mathbb{E}\left[K_h(Y - \boldsymbol{X}^T\beta)X_jX_k\right]\big|_{\beta=\hat{\beta}(\hat{\tau}_{\boldsymbol{x}})} \right| \\
&\quad + \sup_{\boldsymbol{x}\in\mathcal{X}_0} \left| \mathbb{E}\left[K_h(Y - \boldsymbol{X}^T\beta \mid \boldsymbol{X}_i)X_jX_k\right]\big|_{\beta=\hat{\beta}(\hat{\tau}_{\boldsymbol{x}})} - \mathbb{E}\left[f(\boldsymbol{X}^T\beta \mid \boldsymbol{X})X_jX_k\right]\big|_{\beta=\hat{\beta}(\hat{\tau}_{\boldsymbol{x}})} \right| \\
&\quad + \sup_{\boldsymbol{x}\in\mathcal{X}_0} \left| \mathbb{E}\left[f(\boldsymbol{X}^T\beta \mid \boldsymbol{X})X_jX_k\right]\big|_{\beta=\hat{\beta}(\hat{\tau}_{\boldsymbol{x}})} - \mathbb{E}\left[f(\boldsymbol{X}^T\beta \mid \boldsymbol{X})X_jX_k\right]\big|_{\beta=\beta(\tau_{\boldsymbol{x}})} \right|.
\end{aligned}$$

It is routine to show that the first and second terms on the right hand side are $O_P(n^{-1/2}h^{-1})$ and $O(h^2)$, respectively; cf. the proof of Lemma 7. By Taylor expansion, the last term can be bounded by $O_P(\|\hat{\beta}(\hat{\tau}_{\boldsymbol{x}}) - \beta(\tau_{\boldsymbol{x}})\|_{\mathcal{X}_0})$. Observe that

$$\begin{aligned}\|\hat{\beta}(\hat{\tau}_{\boldsymbol{x}}) - \beta(\tau_{\boldsymbol{x}})\|_{\mathcal{X}_0} &\leq \|\hat{\beta} - \beta\|_{[\epsilon,1-\epsilon]} + \|\beta(\hat{\tau}_{\boldsymbol{x}}) - \beta(\tau_{\boldsymbol{x}})\|_{\mathcal{X}_0} \\ &\leq O_P(n^{-1/2} + \|\hat{\tau}_{\boldsymbol{x}} - \tau_{\boldsymbol{x}}\|_{\mathcal{X}_0}) = O_P(n^{-1/2}h^{-3/2}\sqrt{\log n} + h^2).\end{aligned}$$

This completes the proof. □

**Lemma 11.** *Under Assumption 1, we have*

$$\|\hat{\Sigma} - \Sigma\|_\infty = O_P(n^{-1/2}h^{-5/2}\sqrt{\log n} + h).$$

*Proof.* For simplicity of notation, let $\hat{J}_{\boldsymbol{x}_k} = \hat{J}(\hat{\tau}_{\boldsymbol{x}_k})$ and $J_{\boldsymbol{x}_k} = J(\tau_{\boldsymbol{x}_k})$. The difference $\hat{\Sigma}_{\ell,k} - \Sigma_{\ell,k}$ can be decomposed as

$$\begin{aligned}&\left[\frac{\hat{s}_{\boldsymbol{x}_k}(\hat{\tau}_{\boldsymbol{x}_k})\hat{s}_{\boldsymbol{x}_\ell}(\hat{\tau}_{\boldsymbol{x}_\ell})}{\hat{s}''_{\boldsymbol{x}_k}(\hat{\tau}_{\boldsymbol{x}_k})\hat{s}''_{\boldsymbol{x}_\ell}(\hat{\tau}_{\boldsymbol{x}_\ell})}\right] \mathbb{E}_{|\mathcal{D}_n}\left[\frac{1}{h}K'\left(\frac{\hat{\tau}_{\boldsymbol{x}_k} - U}{h}\right)K'\left(\frac{\hat{\tau}_{\boldsymbol{x}_\ell} - U}{h}\right)\right] \boldsymbol{x}_k^T \hat{J}_{\boldsymbol{x}_k}^{-1}\left[\frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i\boldsymbol{X}_i^T\right]\hat{J}_{\boldsymbol{x}_\ell}^{-1}\boldsymbol{x}_\ell \\ &- \left[\frac{s_{\boldsymbol{x}_k}(\tau_{\boldsymbol{x}_k})s_{\boldsymbol{x}_\ell}(\tau_{\boldsymbol{x}_\ell})}{s''_{\boldsymbol{x}_k}(\tau_{\boldsymbol{x}_k})s''_{\boldsymbol{x}_\ell}(\tau_{\boldsymbol{x}_\ell})}\right] \mathbb{E}\left[\frac{1}{h}K'\left(\frac{\tau_{\boldsymbol{x}_k} - U}{h}\right)K'\left(\frac{\tau_{\boldsymbol{x}_\ell} - U}{h}\right)\right] \boldsymbol{x}_k^T J_{\boldsymbol{x}_k}^{-1}\mathbb{E}[\boldsymbol{X}\boldsymbol{X}^T]J_{\boldsymbol{x}_\ell}^{-1}\boldsymbol{x}_\ell.\end{aligned}$$

Observe that

$$\begin{aligned}&\max_{1\leq k,\ell\leq L}\left|\frac{\hat{s}_{\boldsymbol{x}_k}(\hat{\tau}_{\boldsymbol{x}_k})\hat{s}_{\boldsymbol{x}_\ell}(\hat{\tau}_{\boldsymbol{x}_\ell})}{\hat{s}''_{\boldsymbol{x}_k}(\hat{\tau}_{\boldsymbol{x}_k})\hat{s}''_{\boldsymbol{x}_\ell}(\hat{\tau}_{\boldsymbol{x}_\ell})} - \frac{s_{\boldsymbol{x}_k}(\tau_{\boldsymbol{x}_k})s_{\boldsymbol{x}_\ell}(\tau_{\boldsymbol{x}_\ell})}{s''_{\boldsymbol{x}_k}(\tau_{\boldsymbol{x}_k})s''_{\boldsymbol{x}_\ell}(\tau_{\boldsymbol{x}_\ell})}\right| \\ &\quad\leq O_P\left(\|\hat{s}_{\boldsymbol{x}}(\hat{\tau}_{\boldsymbol{x}}) - s_{\boldsymbol{x}}(\tau_{\boldsymbol{x}})\|_{\mathcal{X}_0} \bigvee \|\hat{s}''_{\boldsymbol{x}}(\hat{\tau}_{\boldsymbol{x}}) - s''_{\boldsymbol{x}}(\tau_{\boldsymbol{x}})\|_{\mathcal{X}_0}\right), \quad \text{and} \\ &\|\hat{s}^{(r)}_{\boldsymbol{x}}(\hat{\tau}_{\boldsymbol{x}}) - s^{(r)}_{\boldsymbol{x}}(\tau_{\boldsymbol{x}})\|_{\mathcal{X}_0} \leq \|\hat{s}^{(r)}_{\boldsymbol{x}}(\tau) - s^{(r)}_{\boldsymbol{x}}(\tau)\|_{[\epsilon,1-\epsilon]\times\mathcal{X}_0} + \underbrace{\|s^{(r)}_{\boldsymbol{x}}(\hat{\tau}_{\boldsymbol{x}}) - s^{(r)}_{\boldsymbol{x}}(\tau_{\boldsymbol{x}})\|_{\mathcal{X}_0}}_{=O(\|\hat{\tau}_{\boldsymbol{x}} - \tau_{\boldsymbol{x}}\|_{\mathcal{X}_0})} \\ &\qquad = O_P(n^{-1/2}h^{-5/2}\sqrt{\log n} + h) \quad \text{for } r = 0, 2,\end{aligned}$$

where we have used Lemma 7 in the last line.

Next, we note that

$$\begin{aligned}
&\max_{1\le k,\ell\le L}\left|\boldsymbol{x}_k^T\hat{J}_{\boldsymbol{x}_k}^{-1}\left[\frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i\boldsymbol{X}_i^T\right]\hat{J}_{\boldsymbol{x}_\ell}^{-1}\boldsymbol{x}_\ell - \boldsymbol{x}_k^T J_{\boldsymbol{x}_k}^{-1}\mathbb{E}\left[\boldsymbol{X}_i\boldsymbol{X}_i^T\right]J_{\boldsymbol{x}_\ell}^{-1}\boldsymbol{x}_\ell\right| \\
&\le O_P\left(\max_{1\le k\le L}\|\hat{J}_{\boldsymbol{x}_k}^{-1} - J_{\boldsymbol{x}_k}^{-1}\|_{\mathrm{op}}\bigvee\left\|\frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i\boldsymbol{X}_i^T - \mathbb{E}\left[\boldsymbol{X}\boldsymbol{X}^T\right]\right\|_{\mathrm{op}}\right) \\
&= O_P(n^{-1/2}h^{-3/2}\sqrt{\log n} + h^2),
\end{aligned}$$

where we have used Lemma 10 in the last line.

Finally, observe that

$$\begin{aligned}
&\left|\mathbb{E}_{|\mathcal{D}_n}\left[\frac{1}{h}K'\left(\frac{\hat{\tau}_{\boldsymbol{x}_k}-U}{h}\right)K'\left(\frac{\hat{\tau}_{\boldsymbol{x}_\ell}-U}{h}\right)\right] - \mathbb{E}_{|\mathcal{D}_n}\left[\frac{1}{h}K'\left(\frac{\tau_{\boldsymbol{x}_k}-U}{h}\right)K'\left(\frac{\hat{\tau}_{\boldsymbol{x}_\ell}-U}{h}\right)\right]\right| \\
&\le \|K''\|_\infty h^{-1}|\hat{\tau}_{\boldsymbol{x}_k} - \tau_{\boldsymbol{x}_k}|\mathbb{E}_{|\mathcal{D}_n}\left[\left|\frac{1}{h}K'\left(\frac{\hat{\tau}_{\boldsymbol{x}_\ell}-U}{h}\right)\right|\right] = O_P(n^{-1/2}h^{-5/2}\sqrt{\log n} + h)
\end{aligned}$$

uniformly in $1 \le k, \ell \le L$. Likewise, we have

$$\begin{aligned}
&\left|\mathbb{E}_{|\mathcal{D}_n}\left[\frac{1}{h}K'\left(\frac{\tau_{\boldsymbol{x}_k}-U}{h}\right)K'\left(\frac{\hat{\tau}_{\boldsymbol{x}_\ell}-U}{h}\right)\right] - \mathbb{E}_{|\mathcal{D}_n}\left[\frac{1}{h}K'\left(\frac{\tau_{\boldsymbol{x}_k}-U}{h}\right)K'\left(\frac{\tau_{\boldsymbol{x}_\ell}-U}{h}\right)\right]\right| \\
&\quad = O_P(n^{-1/2}h^{-5/2}\sqrt{\log n} + h)
\end{aligned}$$

uniformly in $1 \le k, \ell \le L$. Combining these estimates, we obtain the conclusion of the lemma. □

**Lemma 12.** *Under Assumptions 1 and 2, we have*

$$\max_{1\le k,\ell\le M}|D_k^T(\hat{\Sigma} - \Sigma)D_\ell| = O_P(n^{-1/2}h^{-5/2}\sqrt{\log n} + h).$$

*Proof.* This follows from the observation that

$$\begin{aligned}
\max_{1\le k,\ell\le M}|D_k^T(\hat{\Sigma} - \Sigma)D_\ell| &= \max_{1\le k,\ell\le M}\left|\sum_{k'\in S_k}\sum_{\ell'\in S_\ell} D_{k,k'}(\hat{\Sigma}_{k',\ell'} - \Sigma_{k',\ell'})D_{\ell,\ell'}\right| \\
&\le \max_{1\le k\le M}|S_k|^2\|D\|_\infty\|\hat{\Sigma} - \Sigma\|_\infty = O_P(n^{-1/2}h^{-5/2}\sqrt{\log n} + h).
\end{aligned}$$

□

We are now in position to prove Theorem 2.

*Proof of Theorem 2.* Let $\hat{G}$ be an $L$-dimensional random vector such that conditionally on $\mathcal{D}_n$, $\hat{G} \sim N(0, \hat{\Sigma})$. We begin with noting that

$$\sup_{b\in\mathbb{R}^M} \left|\mathbb{P}_{|\mathcal{D}_n}\left(n^{-1/2}\textstyle\sum_{i=1}^n \hat{A}\hat{\Psi}_i \le b\right) - \mathbb{P}\left(AG \le b\right)\right| \le \sup_{b\in\mathbb{R}^M} \left|\mathbb{P}_{|\mathcal{D}_n}\left(n^{-1/2}\textstyle\sum_{i=1}^n \hat{A}\hat{\Psi}_i \le b\right) - \right.$$
$$\left. \mathbb{P}_{|\mathcal{D}_n}\left(\hat{A}\hat{G} \le b\right)\right| + \sup_{b\in\mathbb{R}^M} \left|\mathbb{P}_{|\mathcal{D}_n}(\hat{A}\hat{G} \le b) - \mathbb{P}_{|\mathcal{D}_n}(A\hat{G} \le b)\right| + \sup_{b\in\mathbb{R}^M} \left|\mathbb{P}_{|\mathcal{D}_n}(A\hat{G} \le b) - \mathbb{P}(AG \le b)\right|$$
$$:= I + II + III.$$

We first analyze $II$ and $III$. In view of the Gaussian comparison inequality (cf. Lemma 4), to show that $II \vee III = o_P(1)$, it suffices to verify that

$$\left[\|\hat{A}\hat{\Sigma}\hat{A}^T - A\hat{\Sigma}A^T\|_\infty \vee \|A\hat{\Sigma}A^T - A\Sigma A^T\|_\infty\right] \log^2 M = o_P(1). \tag{B.4}$$

Indeed, by Lemma 12 and the assumption (i) of the theorem, we deduce that the bracket on the left is $O_P(n^{-1/2}h^{-5/2}\sqrt{\log n} + h)$. Thus, (B.4) holds under our assumption.

To show that $I = o_P(1)$, we apply Proposition 2.1 in Chernozhukov et al. (2017a) conditionally on $\mathcal{D}_n$ (recall that conditionally on $\mathcal{D}_n$, the vectors $\hat{\Psi}_1, \ldots, \hat{\Psi}_n$ are independent with mean zero). By construction, $n^{-1}\sum_{i=1}^n \mathbb{E}_{|\mathcal{D}_n}[(\hat{A}_k^T\hat{\Psi}_i)^2] = \hat{A}_k^T\hat{\Sigma}\hat{A}_k = D_k^T\hat{\Sigma}D_k/\hat{\Gamma}_k^2$ is bounded way from $0$ uniformly in $k$ with probability approaching one. Similarly to the proof of Theorem 1, we can verify that $\max_{1\le k\le M} n^{-1}\sum_{i=1}^n \mathbb{E}_{|\mathcal{D}_n}[|\hat{A}_k^T\hat{\Psi}_i|^{2+r}] = O_P(h^{-r/2})$ for $r = 1, 2$. Finally,

$$\max_{1\le i\le n} \mathbb{E}_{|\mathcal{D}_n}\left[\max_{1\le k\le M} |\hat{A}_k^T\hat{\Psi}_i|^q\right] \le O_P(h^{-q/2}) \max_{1\le i\le n} \|\boldsymbol{X}_i\|^q = O_P(nh^{-q/2}).$$

Hence, applying Proposition 2.1 in Chernozhukov et al. (2017a), we see that $I = o_P(1)$ as soon as

$$\frac{\log^7(Mn)}{n^{1-2/q}h} \bigvee \frac{\log^3(Mn)}{n^{1-4/q}h} \to 0,$$

but this is satisfied under our assumption. This completes the proof. □

*Proof of Proposition 2.* Theorem 1 implies that

$$\sup_{b\in\mathbb{R}^M}\left|\mathbb{P}\left(D\sqrt{nh^3}(\hat{m}(\boldsymbol{x}_\ell)-m(\boldsymbol{x}_\ell))_{\ell=1}^L\le b\right)-\mathbb{P}(DG\le b)\right|\to 0.$$

Since the variances of the coordinates of $G$ are bounded, we see that $\mathbb{E}[\|G\|_\infty]=O(\sqrt{\log M})$ by Lemma 2.2.2 in van der Vaart and Wellner (1996). Hence, we have

$$\left\|D\sqrt{nh^3}(\hat{m}(\boldsymbol{x}_\ell)-m(\boldsymbol{x}_\ell))_{\ell=1}^L\right\|_\infty=O_P(\sqrt{\log M}).$$

Combining Condition (i) in the statement of Theorem 2, we see that

$$\left\|(\hat{\Gamma}^{-1}-\Gamma^{-1})D\sqrt{nh^3}(\hat{m}(\boldsymbol{x}_\ell)-m(\boldsymbol{x}_\ell))_{\ell=1}^L\right\|_\infty=o_P(1/\sqrt{\log M}).$$

The rest of the proof is analogous to the last part of Theorem 1. We omit the details for brevity. □

**Proof of Theorem 3**

We start with proving the following uniform Bahadur representation for the quantile regression estimator $\hat{\beta}^*$ based on the nonparametric bootstrap samples $(Y_i^*,\boldsymbol{X}_i^*)_{i=1}^n$. We define $U_i^*=F(Y_i^*\mid\boldsymbol{X}_i^*)$ where $F(y\mid\boldsymbol{x})$ is the conditional distribution function of $Y$ given $\boldsymbol{X}$ (see Remark 12).

**Lemma 13.** *Suppose Assumption 1 holds. Then we have, for arbitrarily small $\gamma>0$,*

$$\hat{\beta}^*(\tau)-\beta(\tau)=J(\tau)^{-1}\left[\frac{1}{n}\sum_{i=1}^n\{\tau-I(U_i^*\le\tau)\}\boldsymbol{X}_i^*\right]+O_P(n^{-3/4+\gamma}),\qquad\text{(B.5)}$$

*uniformly in $\tau\in[\epsilon/2,1-\epsilon/2]$.*

*Proof.* We will prove the following equivalent form of (B.5),

$$\hat{\beta}^*(\tau) - \beta(\tau) = J(\tau)^{-1}\left[\frac{1}{n}\sum_{i=1}^{n}\pi_i\{\tau - I(U_i \le \tau)\}\boldsymbol{X}_i\right] + O_P(n^{-3/4+\gamma}), \qquad \text{(B.6)}$$

where $(\pi_1, \ldots, \pi_n)$ is a multinomial random vector with parameters $n$ and (probabilities) $1/n, \ldots, 1/n$. We will divide the proof into two steps. In the following proof, $C$ is a generic constant independent of $n$ whose value may vary from line to line.

**Step 1**. In this step, we will show that $\sup_{\tau\in[\epsilon/2,1-\epsilon/2]} \|\hat{\beta}^*(\tau) - \beta(\tau)\| = O_P(n^{-1/2})$. To this end, we introduce the following quantities

$$\begin{aligned}
R_n &:= \Big\{(\tau,\beta) \in \mathcal{U} \times \mathbb{R}^d : \|\beta - \beta(\tau)\| \le r_n\Big\},\\
\Upsilon_0 &:= \sup_{\tau\in\mathcal{U}}\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\pi_i\{\tau - I(Y_i \le \boldsymbol{X}_i^T\beta(\tau))\}\boldsymbol{X}_i\right\|,\\
\Upsilon_1 &:= \sup_{(\tau,\beta)\in R_n}\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\big\{\pi_i\{I(Y_i \le \boldsymbol{X}_i^T\beta) - I(Y_i \le \boldsymbol{X}_i^T\beta(\tau))\}\boldsymbol{X}_i\right.\\
&\qquad\qquad \left. - \mathbb{E}[\{I(Y \le \boldsymbol{X}^T\beta) - \tau\}\boldsymbol{X}]\big\}\right\|,\\
\Upsilon_2 &:= \sup_{(\tau,\beta)\in R_n} n^{1/2}\Big\|\mathbb{E}[\{\tau - I(Y \le \boldsymbol{X}^T\beta)\}\boldsymbol{X}] - J(\tau)(\beta - \beta(\tau))\Big\|.
\end{aligned}$$

where we define $\mathcal{U} := [\epsilon/2, 1-\epsilon/2]$ and $(r_n)_{n=1}^{\infty}$ is a sequence of constants to be specified.

In view of the proof of Lemma 3 in Belloni et al. (2019a), it suffices to show that $\Upsilon_0 = O_P(1)$, $\Upsilon_1 = o_P(1)$ and $\Upsilon_2 = o_P(1)$ when $r_n = O(n^{-1/2})$. We shall bound the three terms in what follows.

The term $\Upsilon_0$ is the supremum of a multiplier empirical process, and we will apply a multiplier inequality developed in Han and Wellner (2019). Define the function class

$$\mathcal{F}_1 := \Big\{(y,\boldsymbol{x}) \mapsto (\tau - I\{y \le \boldsymbol{x}^T\beta(\tau)\})\alpha^T\boldsymbol{x} : \tau \in \mathcal{U}, \alpha \in \mathbb{S}^{d-1}\Big\},$$

where $\mathbb{S}^{d-1} = \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\| = 1\}$. Then $\Upsilon_0 = \|n^{-1/2}\sum_{i=1}^n \pi_i f(\boldsymbol{X}_i, Y_i)\|_{\mathcal{F}_1}$. We first verify that

$$\mathbb{E}\big[\|\mathbb{G}_n\|_{\mathcal{F}_1}\big] \leq Cd^{1/2}. \tag{B.7}$$

This can be proved as follows

$$\begin{aligned}\mathbb{E}\big[\|\mathbb{G}_n\|_{\mathcal{F}_1}\big] &\leq \mathbb{E}\bigg[\sup_{\tau\in\mathcal{U}}\Big[\sum_{j=1}^d \big(\mathbb{G}_n\big[\{\tau - I(Y_i \leq \boldsymbol{X}_i^T\beta(\tau))\}X_{ij}\big]\big)^2\Big]^{1/2}\bigg] \\ &\leq \bigg[\sum_{j=1}^d \mathbb{E}\Big[\sup_{\tau\in\mathcal{U}}\big(\mathbb{G}_n\big[\{\tau - I(U_i \leq \tau)\}X_{ij}\big]\big)^2\Big]\bigg]^{1/2}.\end{aligned}$$

For any $1 \leq j \leq d$, by Theorem 2.14.1 in van der Vaart and Wellner (1996),

$$\mathbb{E}\Big[\sup_{\tau\in\mathcal{U}}\big(\mathbb{G}_n\big[\{\tau - I(U_i \leq \tau)\}X_{ij}\big]\big)^2\Big] \leq C\mathbb{E}\left[|X_{1j}|^2\right] = O(1).$$

This leads to (B.7).

Now, by Lemma 2.3.6 in van der Vaart and Wellner (1996), (B.7) implies the following bound for the symmetrized empirical process

$$\mathbb{E}\Big[\Big\|\sum_{i=1}^n \xi_i f(\boldsymbol{X}_i, Y_i)\Big\|_{\mathcal{F}_1}\Big] \leq Cd^{1/2}n^{1/2},$$

where $\xi_i$ ($1 \leq i \leq n$) are i.i.d. Rademacher random variables independent of the data. Given the above bound, we can apply Corollary 1 in Han and Wellner (2019) (recall $d$ is fixed) to conclude that

$$\mathbb{E}[\Upsilon_0] \leq Cd^{1/2}\|\pi_1\|_{2,1} = O(d^{1/2}).$$

This implies that $\Upsilon_0 = O_P(1)$.

We can bound $\Upsilon_1$ similarly to $\Upsilon_0$. Define the function class

$$\begin{aligned}\mathcal{F}_2 = \Big\{(y, \boldsymbol{x}) \mapsto \{I(y \leq \boldsymbol{x}^T\beta) - I(y \leq \boldsymbol{x}^T\beta(\tau))\}\alpha^T\boldsymbol{x} - \mathbb{E}\big[\{I(Y \leq \boldsymbol{X}^T\beta) - \tau\}\alpha^T\boldsymbol{X}\big] \\ : (\tau, \beta) \in R_n, \alpha \in \mathbb{S}^{d-1}\Big\},\end{aligned}$$

so that $\Upsilon_1 = \|n^{-1/2}\sum_{i=1}^n \pi_i f(\boldsymbol{X}_i, Y_i)\|_{\mathcal{F}_2}$. Applying Lemma 2, we have

$$\mathbb{E}\Big[\Big\|\sum_{i=1}^n f(\boldsymbol{X}_i, Y_i)\Big\|_{\mathcal{F}_2}\Big] = o(n^{1/4}\log n).$$

Now we apply Corollary 1 in Han and Wellner (2019) (with $\Psi_n(t) = Ct^{1/4+\gamma}$ for some arbitrary small $\gamma > 0$ in the proof there) to conclude that

$$\mathbb{E}[\Upsilon_1] \le Cn^{-1/4+\gamma},$$

which implies that $\Upsilon_1 = O_P(n^{-1/4+\gamma}) = o_P(1)$.

For $\Upsilon_2$, we proceed as in Lemma 33 of Belloni et al. (2019a) to see that

$$\Upsilon_2 \le \sup_{\substack{(\tau,\beta)\in R_n \\ \alpha\in\mathbb{S}^{d-1}}} \sqrt{n}\Big|\mathbb{E}\big[\big(f(\boldsymbol{X}^T\tilde{\beta}(\tau) \mid \boldsymbol{X}) - f(\boldsymbol{X}\beta(\tau) \mid \boldsymbol{X})\big)\cdot(\alpha^T\boldsymbol{X})\cdot\big(\boldsymbol{X}^T(\beta-\beta(\tau))\big)\big]\Big|,$$

where $\tilde{\beta}(\tau)$ lies on the line segment between $\beta(\tau)$ and $\beta$. We further bound the right-hand side as

$$\begin{aligned}
\Upsilon_2 &\le C\sqrt{n}\sup_{\substack{(\tau,\beta)\in R_n \\ \alpha\in\mathbb{S}^{d-1}}} \mathbb{E}\Big[|\alpha^T\boldsymbol{X}|\big|(\tilde{\beta}(\tau)-\beta(\tau))^T\boldsymbol{X}\boldsymbol{X}^T(\beta-\beta(\tau))\big|\Big] \\
&\le C\sqrt{n}\sup_{\substack{(\tau,\beta)\in R_n \\ \alpha\in\mathbb{S}^{d-1}}} \sqrt{\mathbb{E}\left[\alpha^T\boldsymbol{X}\boldsymbol{X}^T\alpha\right]}\cdot\sqrt{\mathbb{E}\Big[\big|(\tilde{\beta}(\tau)-\beta(\tau))^T\boldsymbol{X}\boldsymbol{X}^T(\beta-\beta(\tau))\big|^2\Big]} \\
&\le C\sqrt{n}\cdot O(1)\cdot r_n^2\sqrt{\mathbb{E}\|\boldsymbol{X}\|^4} = O(n^{-1/2}) = o(1).
\end{aligned}$$

These bounds imply the conclusion of this step, in view of the proof of Lemma 3 in Belloni et al. (2019a).

**Step 2.** We finish the proof of the lemma. Define

$$r(\tau) = J(\tau)\big(\hat{\beta}^*(\tau) - \beta(\tau)\big) - \frac{1}{n}\sum_{i=1}^n \pi_i\{\tau - I(U_i \le \tau)\}\boldsymbol{X}_i.$$

It suffices to show that $\sup_{\tau\in\mathcal{U}}\|r(\tau)\| = O_P(n^{-3/4+\gamma})$ for arbitrarily small $\gamma > 0$. We note that, by Step 1, for any $B_n \to \infty$ arbitrarily slowly, $(\tau, \beta^*(\tau)) \in R_n$ with

$r_n = B_n n^{-1/2}$ holds with probability approaching one. Hence, taking such $R_n$, with probability $1 - o(1)$,

$$\sup_{\tau \in \mathcal{U}} \|r(\tau)\| \leq n^{-1/2}(\Upsilon_1 + \Upsilon_2 + \Upsilon_3),$$

where $\Upsilon_3 := \sup_{\tau \in \mathcal{U}} \|n^{-1/2} \sum_{i=1}^n \pi_i \{\tau - I(Y_i \leq \boldsymbol{X}_i^T \hat{\beta}^*(\tau))\} \boldsymbol{X}_i\|$. By Step 1, taking $B_n \to \infty$ sufficiently slowly, we have

$$\Upsilon_1 = O_P(n^{-1/4+\gamma}) \quad \text{and} \quad \Upsilon_2 = O_P(n^{-1/2+\gamma}).$$

To bound $\Upsilon_3$, from the proof of Lemma 34 in Belloni et al. (2019a),we can deduce that

$$\Upsilon_3 \leq \frac{d}{\sqrt{n}} \max_{1 \leq i \leq n} \|\pi_i \boldsymbol{X}_i\|.$$

We further bound the right-hand side as

$$\begin{aligned}
\Upsilon_3 &\leq \frac{d}{\sqrt{n}} \cdot \max_{1 \leq i \leq n} |\pi_i| \cdot \max_{1 \leq i \leq n} \|\boldsymbol{X}_i\| \\
&= \frac{d}{\sqrt{n}} \cdot o_P\Big(\frac{\log n}{\log \log n}\Big) \cdot o_P(n^{1/q}) \qquad \text{(Section 4 of Raab and Steger (1998))} \\
&= o_P(d n^{1/q - 1/2} \log n / \log \log n).
\end{aligned}$$

Hence we have shown that

$$\begin{aligned}
\sup_{\tau \in \mathcal{U}} \|r(\tau)\| &= O_P(n^{-3/4+\gamma} + n^{-1+\gamma} + d n^{1/q-1} \log n / \log \log n) \\
&= O_P(n^{-3/4+\gamma}),
\end{aligned}$$

which finishes the proof. □

Define $\bar{\Psi} := n^{-1} \sum_{i=1}^n \Psi_i$ and $\tilde{\Sigma} := n^{-1} \sum_{i=1}^n (\Psi_i - \bar{\Psi})(\Psi_i - \bar{\Psi})^T$.

**Lemma 14.** *Under Assumption 1, we have*

$$\|\tilde{\Sigma} - \Sigma\|_\infty = O_P(n^{-1/2} h^{-1/2} \sqrt{\log n} + n^{2/q-1} h^{-1} \log n).$$

*Proof.* Recall that $J_{\boldsymbol{x}} = J(\tau_{\boldsymbol{x}})$. The difference $\tilde{\Sigma}_{j,k} - \Sigma_{j,k}$ can be decomposed as

$$\begin{aligned}
&\left[\frac{s_{\boldsymbol{x}_k}(\tau_{\boldsymbol{x}_k})s_{\boldsymbol{x}_j}(\tau_{\boldsymbol{x}_j})}{s''_{\boldsymbol{x}_k}(\tau_{\boldsymbol{x}_k})s''_{\boldsymbol{x}_j}(\tau_{\boldsymbol{x}_j})}\right]\left\{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K'\left(\frac{\tau_{\boldsymbol{x}_k}-U_i}{h}\right)K'\left(\frac{\tau_{\boldsymbol{x}_j}-U_i}{h}\right)\boldsymbol{x}_k^T J_{\boldsymbol{x}_k}^{-1}\boldsymbol{X}_i\boldsymbol{X}_i^T J_{\boldsymbol{x}_j}^{-1}\boldsymbol{x}_j\right.\\
&\quad-\mathbb{E}\left[\frac{1}{h}K'\left(\frac{\tau_{\boldsymbol{x}_k}-U}{h}\right)K'\left(\frac{\tau_{\boldsymbol{x}_j}-U}{h}\right)\right]\boldsymbol{x}_k^T J_{\boldsymbol{x}_k}^{-1}\mathbb{E}[\boldsymbol{X}\boldsymbol{X}^T]J_{\boldsymbol{x}_j}^{-1}\boldsymbol{x}_j\\
&\quad\left.-\frac{1}{h}\left[\sum_{i=1}^{n}K'\left(\frac{\tau_{\boldsymbol{x}_k}-U_i}{h}\right)\boldsymbol{x}_k^T J_{\boldsymbol{x}_k}^{-1}\boldsymbol{X}_i\right]\left[\sum_{i=1}^{n}K'\left(\frac{\tau_{\boldsymbol{x}_j}-U_i}{h}\right)\boldsymbol{x}_j^T J_{\boldsymbol{x}_j}^{-1}\boldsymbol{X}_i\right]\right\}.
\end{aligned}$$

We define

$$\begin{aligned}
I_{jk} &:= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K'\left(\frac{\tau_{\boldsymbol{x}_k}-U_i}{h}\right)K'\left(\frac{\tau_{\boldsymbol{x}_j}-U_i}{h}\right)\boldsymbol{x}_k^T J_{\boldsymbol{x}_k}^{-1}\boldsymbol{X}_i\boldsymbol{X}_i^T J_{\boldsymbol{x}_j}^{-1}\boldsymbol{x}_j\\
&\quad-\mathbb{E}\left[\frac{1}{h}K'\left(\frac{\tau_{\boldsymbol{x}_k}-U}{h}\right)K'\left(\frac{\tau_{\boldsymbol{x}_j}-U}{h}\right)\right]\boldsymbol{x}_k^T J_{\boldsymbol{x}_k}^{-1}\mathbb{E}[\boldsymbol{X}\boldsymbol{X}^T]J_{\boldsymbol{x}_j}^{-1}\boldsymbol{x}_j;\\
II_{jk} &:= \frac{1}{h}\left[\sum_{i=1}^{n}K'\left(\frac{\tau_{\boldsymbol{x}_k}-U_i}{h}\right)\boldsymbol{x}_k^T J_{\boldsymbol{x}_k}^{-1}\boldsymbol{X}_i\right]\left[\sum_{i=1}^{n}K'\left(\frac{\tau_{\boldsymbol{x}_j}-U_i}{h}\right)\boldsymbol{x}_j^T J_{\boldsymbol{x}_j}^{-1}\boldsymbol{X}_i\right].
\end{aligned}$$

We first bound $I_{jk}$. Define the function class

$$\begin{aligned}
\mathcal{F}_1 := \Big\{&(u,\boldsymbol{x})\mapsto h^{-1}K'((\tau_1-u)/h)K'((\tau_2-u)/h)(\boldsymbol{x}_1^T J_{\boldsymbol{x}_1}^{-1}\boldsymbol{x})(\boldsymbol{x}^T J_{\boldsymbol{x}_2}^{-1}\boldsymbol{x}_2):\\
&\tau_1,\tau_2\in[\epsilon,1-\epsilon],\ \boldsymbol{x}_1,\boldsymbol{x}_2\in\mathcal{X}_0\Big\}.
\end{aligned}$$

Then we have $\max_{1\le j,k\le L}|I_{jk}| \le \sup_{f\in\mathcal{F}_1}|n^{-1}\sum_{i=1}^{n}\{f(U_i,\boldsymbol{X}_i)-\mathbb{E}[f(U,\boldsymbol{X})]\}|$. Applying Lemma 2, we have

$$\sup_{f\in\mathcal{F}_1}\left|n^{-1}\sum_{i=1}^{n}\{f(U_i,\boldsymbol{X}_i)-\mathbb{E}[f(U,\boldsymbol{X})]\}\right| = O_P\big(n^{-1/2}h^{-1/2}\sqrt{\log n}+n^{2/q-1}h^{-1}\log n\big).$$

Hence we have shown that

$$\max_{1\le j,k\le L}|I_{jk}| = O_P\big(n^{-1/2}h^{-1/2}\sqrt{\log n}+n^{2/q-1}h^{-1}\log n\big).$$

For $II_{jk}$, we define the following function class

$$\mathcal{F}_2 := \Big\{(u,\boldsymbol{x})\mapsto h^{-1/2}K'((\tau-u)/h)\boldsymbol{x}_0^T J_{\boldsymbol{x}_0}^{-1}\boldsymbol{x}:\tau\in[\epsilon,1-\epsilon],\boldsymbol{x}_0\in\mathcal{X}_0\Big\}.$$

Then we have

$$\max_{1\leq j,k\leq L}|II_{jk}| \leq \sup_{f\in\mathcal{F}_2}\left\{n^{-1}\sum_{i=1}^{n}\{f(U_i,\boldsymbol{X}_i)-\mathbb{E}[f(U,\boldsymbol{X})]\}\right\}^2.$$

Similarly to the previous case, by using Lemma 2, we can show that

$$\sup_{f\in\mathcal{F}_2}\left|n^{-1}\sum_{i=1}^{n}\{f(U_i,\boldsymbol{X}_i)-\mathbb{E}[f(U,\boldsymbol{X})]\}\right| = O_P(n^{-1/2}\sqrt{\log n}),$$

which implies that $\max_{1\leq j,k\leq L}|II_{jk}| = O_P(n^{-1}\log n)$. Combining the above bounds, we obtain the desired result. □

Using Lemma 14, we have the following result; cf. the proof of Lemma 12.

**Lemma 15.** *Under Assumptions 1 and 2, we have*

$$\max_{1\leq k,\ell\leq M}|D_k^T(\tilde{\Sigma}-\Sigma)D_\ell| = O_P(n^{-1/2}h^{-1/2}\sqrt{\log n}+n^{2/q-1}h^{-1}\log n).$$

We are now in position to prove Theorem 3.

*Proof of Theorem 3.* We begin with noting that, using the Bahadur representation in Lemma 13,we can establish the following asymptotic linear representation for the bootstrap mode estimator by a similar analysis to the proof of Theorem 1 coupled with the multiplier inequality techniques as in the proof of Lemma 13:

$$\begin{aligned}(\hat{m}^*(\boldsymbol{x}_\ell)-m(\boldsymbol{x}_\ell))_{\ell=1}^L &= (nh^{3/2})^{-1}\textstyle\sum_{i=1}^n\Psi_i^*+R_n^*,\\ \sqrt{nh^3}\|R_n^*\|_\infty &= O_P(h^{1/2}+n^{-1/2+\gamma}h^{-5/2}+n^{1/2}h^{7/2}),\end{aligned} \tag{B.8}$$

where $\Psi_i^* = (\psi_{\boldsymbol{x}_1}(U_i^*,\boldsymbol{X}_i^*),\cdots,\psi_{\boldsymbol{x}_L}(U_i^*,\boldsymbol{X}_i^*))^T$. Hence we have

$$\begin{aligned}\sqrt{nh^3}\hat{A}(\hat{m}^*(\boldsymbol{x}_\ell)-\hat{m}(\boldsymbol{x}_\ell))_{\ell=1}^L &= n^{-1/2}\textstyle\sum_{i=1}^n\hat{A}(\Psi_i^*-\Psi_i)+\sqrt{nh^3}\hat{A}(R_n^*-R_n)\\ &:= n^{-1/2}\textstyle\sum_{i=1}^n(\hat{A}\Psi_i^*-\hat{A}\bar{\Psi})+\sqrt{nh^3}\hat{A}\tilde{R}_n.\end{aligned}$$

Now, we divide the rest of the proof into two steps.

**Step 1.** We will show that

$$\sup_{b\in\mathbb{R}^M}\left|\mathbb{P}_{|\mathcal{D}_n}\left(n^{-1/2}\textstyle\sum_{i=1}^n(\hat{A}\Psi_i^*-\hat{A}\bar{\Psi})\le b\right)-\mathbb{P}\left(AG\le b\right)\right|\xrightarrow{P}0. \tag{B.9}$$

We note that

$$\begin{aligned}
&\sup_{b\in\mathbb{R}^M}\left|\mathbb{P}_{|\mathcal{D}_n}\left(n^{-1/2}\textstyle\sum_{i=1}^n(\hat{A}\Psi_i^*-\hat{A}\bar{\Psi})\le b\right)-\mathbb{P}\left(AG\le b\right)\right|\\
&\le \sup_{b\in\mathbb{R}^M}\left|\mathbb{P}_{|\mathcal{D}_n}\left(n^{-1/2}\textstyle\sum_{i=1}^n(\hat{A}\Psi_i^*-\hat{A}\bar{\Psi})\le b\right)-\mathbb{P}_{|\mathcal{D}_n}\left(\hat{A}\tilde{G}\le b\right)\right|\\
&+\sup_{b\in\mathbb{R}^M}\left|\mathbb{P}_{|\mathcal{D}_n}\left(\hat{A}\tilde{G}\le b\right)-\mathbb{P}_{|\mathcal{D}_n}\left(A\tilde{G}\le b\right)\right|+\sup_{b\in\mathbb{R}^M}\left|\mathbb{P}_{|\mathcal{D}_n}\left(A\tilde{G}\le b\right)-\mathbb{P}\left(AG\le b\right)\right|\\
&:= I+II+III,
\end{aligned}$$

where $\tilde{G}\sim N(0,\tilde{\Sigma})$ and recall $\tilde{\Sigma}:=n^{-1}\sum_{i=1}^n(\Psi_i-\bar{\Psi})(\Psi_i-\bar{\Psi})^T$.

We first analyze $II$ and $III$. In view of the Gaussian comparison inequality (cf. Lemma 4), to show that $II\vee III=o_P(1)$, it suffices to verify that

$$\left[\|\hat{A}\tilde{\Sigma}\hat{A}^T-A\tilde{\Sigma}A^T\|_\infty\vee\|A\tilde{\Sigma}A^T-A\Sigma A^T\|_\infty\right]\log^2 M=o_P(1). \tag{B.10}$$

Indeed, by Lemma 15 and Condition (i) of the theorem, we can deduce that the bracket on the left hand side is $O_P(n^{-1/2}h^{-5/2}\sqrt{\log n}+h)$. Thus, (B.10) holds under our assumption.

To show that $I=o_P(1)$, we apply Proposition 2.1 in Chernozhukov et al. (2017a) conditionally on $\mathcal{D}_n$ (recall that conditionally on $\mathcal{D}_n$, the vectors $\Psi_1^*-\bar{\Psi},\ldots,\Psi_n^*-\bar{\Psi}$ are independent with mean zero). By construction, $n^{-1}\sum_{i=1}^n\mathbb{E}_{|\mathcal{D}_n}[(\hat{A}_k^T\Psi_i^*-\hat{A}_k^T\bar{\Psi})^2]=\hat{A}_k^T\tilde{\Sigma}\hat{A}_k=D_k^T\tilde{\Sigma}D_k/\hat{\Gamma}_k^2$ is bounded away from zero uniformly over $1\le k\le M$ with probability approaching one. Similarly to the proof of Theorem 1, we can verify that $\max_{1\le k\le M}n^{-1}\sum_{i=1}^n\mathbb{E}_{|\mathcal{D}_n}[|\hat{A}_k^T\Psi_i^*-$

$\hat{A}_k^T\bar{\Psi}|^{2+r}] = O_P(h^{-r/2})$ for $r = 1, 2$. Finally,

$$\max_{1\le i\le n}\mathbb{E}_{|\mathcal{D}_n}\left[\max_{1\le k\le M}|\hat{A}_k^T\Psi_i^* - \hat{A}_k^T\bar{\Psi}|^q\right] \le O(1)\max_{1\le i\le n}\max_{1\le k\le M}|\hat{A}_k^T\Psi_i|^q$$
$$\le O_P(h^{-q/2})\max_{1\le i\le n}\|\boldsymbol{X}_i\|^q = O_P(nh^{-q/2}).$$

Hence, applying Proposition 2.1 in Chernozhukov et al. (2017a), we see that $I = o_P(1)$ as soon as

$$\frac{\log^7(Mn)}{n^{1-2/q}h}\bigvee\frac{\log^3(Mn)}{n^{1-4/q}h}\to 0,$$

but this is satisfied under our assumption. This completes Step 1.

**Step 2.** We finish the proof by a similar analysis as Step 2 in the proof of Theorem 1. Define $\tilde{\delta}_n = h^{1/2}+n^{-1/2+\gamma}h^{-5/2}+n^{1/2}h^{7/2}$. Combining the analysis before Step 1 and the fact that $\sqrt{nh^3}\|R_n\|_\infty = O_P(h^{1/2}+n^{-1/2}h^{-5/2}\log n + n^{1/2}h^{7/2})$, we have $\sqrt{nh^3}\|\tilde{R}_n\|_\infty = O_P(\tilde{\delta}_n)$. Similarly to Step 2 in the proof of Theorem 1, we can show that $\sqrt{nh^3}\|\hat{A}\tilde{R}_n\|_\infty = O_P(\tilde{\delta}_n)$. The rest of the proof is analogous to the last part of Theorem 1. We omit the details for brevity. □

**Proofs for Appendix B.1**

*Proof of Proposition 5.* Since $K$ is supported in $[-1, 1]$, if $|\tau_{\boldsymbol{x}_k} - \tau_{\boldsymbol{x}_\ell}| > 2h$, then

$$\mathbb{E}\left[K'\left(\frac{\tau_{\boldsymbol{x}_k}-U}{h}\right)K'\left(\frac{\tau_{\boldsymbol{x}_\ell}-U}{h}\right)\right] = 0.$$

Thus, $\Sigma = \mathrm{diag}\{\sigma^2_{\boldsymbol{x}_1},\ldots,\sigma^2_{\boldsymbol{x}_L}\}$, so that Theorem 1 implies that

$$\sup_{b\in\mathbb{R}}\left|\mathbb{P}(\zeta_n\le b) - \mathbb{P}\big(\max_{1\le\ell\le L}|W_\ell|\le b\big)\right|\to 0,$$

where $W_1,\ldots,W_L \sim N(0,1)$ i.i.d. The rest of the proof follows from standard extreme value theory; cf. Theorem 1.5.3 in Leadbetter et al. (1983). □

*Proof of Lemma 1.* Let $q_n(\alpha)$ denote the $\alpha$-quantile of $Z_n$. By assumption, we may choose a sequence $\delta_n \to 0$ such that

$$\sup_{t\in\mathbb{R}} |\mathbb{P}(Y_n \le t) - \mathbb{P}(Z_n \le t)| \le \delta_n \quad \text{and}$$
$$\mathbb{P}\left(\sup_{t\in\mathbb{R}} |\mathbb{P}(W_n \le t \mid \mathcal{C}_n) - \mathbb{P}(Z_n \le t)| > \delta_n\right) \le \delta_n.$$

The latter follows from the fact that the Ky Fan metric metrizes convergence in probability. Define the event $E_n = \{\sup_{t\in\mathbb{R}} |\mathbb{P}(W_n \le t \mid \mathcal{C}_n) - \mathbb{P}(Z_n \le t)| \le \delta_n\}$. On this event,

$$\mathbb{P}(W_n \le q_n(\alpha+\delta_n) \mid \mathcal{C}_n) \ge \underbrace{\mathbb{P}(Z_n \le q_n(\alpha+\delta_n))}_{=\alpha+\delta_n} - \delta_n = \alpha,$$

so that $\hat{q}_n(\alpha) \le q_n(\alpha+\delta_n)$. Thus,

$$\mathbb{P}(Y_n \le \hat{q}_n(\alpha)) \le \mathbb{P}(Y_n \le q_n(\alpha+\delta_n)) + \delta_n \le \mathbb{P}(Z_n \le q_n(\alpha+\delta_n)) + 2\delta_n = \alpha + 3\delta_n.$$

Likewise, on the event $E_n$,

$$\mathbb{P}(Z_n \le t)|_{t=\hat{q}_n(\alpha)} \ge \underbrace{\mathbb{P}(W_n \le t \mid \mathcal{C}_n)|_{t=\hat{q}_n(\alpha)}}_{\ge\alpha} - \delta_n \ge \alpha - \delta_n,$$

so that $\hat{q}_n(\alpha) \ge q_n(\alpha-\delta_n)$. Arguing as in the previous case, we see that $\mathbb{P}(Y_n \le \hat{q}_n(\alpha)) \ge \alpha - 3\delta_n$. This completes the proof.

□

## B.4 Proofs for Section 2.5

Recall that $\mathbb{S}^{d-1}$ is the unit sphere in $\mathbb{R}^d$, i.e., $\mathbb{S}^{d-1} := \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\| = 1\}$. Also recall that in Section 5, we allow $d = d_n \to \infty$.

### B.4.1 Proof of Theorem 4

Overall, the proof is analogous to that of Theorem 1. The following Banadur representation is taken from Belloni et al. (2019a).

**Lemma 16.** *Under Assumption 3, we have*

$$\hat{\beta}(\tau) - \beta(\tau) = J(\tau)^{-1} \left[ \frac{1}{n} \sum_{i=1}^{n} \{\tau - I(U_i \leq \tau))\} \boldsymbol{X}_i \right] + \check{R}_n(\tau)$$

*with* $\|\check{R}_n\|_{[\epsilon/2,1-\epsilon/2]} = O_P(n^{-3/4} d\sqrt{\log n})$ *and* $\|n^{-1} \sum_{i=1}^{n} \{\tau - I(U_i \leq \tau)\} \boldsymbol{X}_i\|_{[\epsilon/2,1-\epsilon/2]} = O_P(\sqrt{d/n})$

*Proof.* See Theorems 1 and 2 in Belloni et al. (2019a). □

The rates of convergence of $\hat{Q}_{\boldsymbol{x}}^{(r)}(\tau_{\boldsymbol{x}})$ change as follows.

**Lemma 17.** *Under the conditions of Theorem 4, we have*

$$\sup_{\substack{\boldsymbol{x} \in \mathcal{X}_0 \\ \tau \in [\epsilon, 1-\epsilon]}} |\hat{Q}_{\boldsymbol{x}}^{(r)}(\tau) - Q_{\boldsymbol{x}}^{(r)}(\tau)| = \begin{cases} O_P\left(n^{-1/2} d + h^2\right) & \text{if } r = 0 \\ O_P\left(n^{-1/2} h^{-r+1/2} d\sqrt{\log n} + h^2\right) & \text{if } r = 1 \text{ or } 2 \\ O_P\left(n^{-1/2} h^{-5/2} d\sqrt{\log n} + h\right) & \text{if } r = 3 \end{cases}$$

*Proof.* We divide the proof into two steps.

**Step 1**. We will show that

$$\sup_{\substack{\boldsymbol{x} \in \mathcal{X}_0 \\ \tau \in [\epsilon, 1-\epsilon]}} \left| \frac{1}{nh^r} \sum_{i=1}^{n} \boldsymbol{x}^T J(\tau)^{-1} \boldsymbol{X}_i \left\{ K^{(r-1)}\left(\frac{\tau - U_i}{h}\right) - hI(r=1) \right\} \right| = O_P\left(n^{-1/2} h^{-r+1/2} d\sqrt{\log n}\right),$$

for $r = 1, 2, 3$. The proof is analogous to that of Lemma 6, so we only point out required modifications. The envelope function $F$ should be modified to

$F(u, \boldsymbol{x}') = C\sqrt{d}\|\boldsymbol{x}'\|$ for some constant $C$, and note that the VC constant $V$ is of order $V = O(d)$. Observe that

$$\sup_{\substack{\boldsymbol{x}\in\mathcal{X}_0 \\ \tau\in[\epsilon,1-\epsilon]}} \mathbb{E}[\{K^{(r-1)}((\tau-U)/h)\boldsymbol{x}^T J(\tau)^{-1}\boldsymbol{X}\}^2] \leq O(d)\int_0^1 K^{(r-1)}((\tau-u)/h)^2 du = O(hd),$$

and $\mathbb{E}[\max_{1\leq i\leq n} F^2(U_i, \boldsymbol{X}_i)] = O(d^2)$ (as $\|\boldsymbol{X}\| \leq C_3\sqrt{d}$). Applying Lemma 2 leads to the above rates.

**Step 2**. We will show the conclusion of the lemma. This part is analogous to the proof of Lemma 7, so we only point out required modifications. The $r = 0$ follows from Lemma 16 and Taylor expansion. For $1 \leq r \leq 3$, combining Lemma 16, change of variables, and Taylor expansion, we can bound $\sup_{\boldsymbol{x}\in\mathcal{X}_0;\tau\in[\epsilon,1-\epsilon]} |\hat{Q}^{(r)}_{\boldsymbol{x}}(\tau) - Q^{(r)}_{\boldsymbol{x}}(\tau)|$ by

$$\sup_{\substack{\boldsymbol{x}\in\mathcal{X}_0 \\ \tau\in[\epsilon,1-\epsilon]}} \left| \frac{1}{nh^r}\sum_{i=1}^n \int \boldsymbol{x}^T J(\tau-th)^{-1}\boldsymbol{X}_i \left\{\tau - th - I\left(U_i \leq \tau - th\right)\right\} K^{(r)}(t)dt \right|$$

$$+ \underbrace{O_P(n^{-3/4}d^{3/2}h^{-r}\sqrt{\log n})}_{=o_P(n^{-1/2}dh^{-r+1/2}\sqrt{\log n})} + O(h^2 I(r=1,2) + hI(r=3)).$$

Replacing $J(\tau - th)$ by $J(\tau)$ in the first term on the right hand side results in an error of order $O_P(n^{-1/2}dh^{-r+1})$. Given Step 1, the rest of the proof is completely analogous to the last part of the proof of Lemma 7. □

**Remark 15** (Expansion of $\hat{Q}''_{\boldsymbol{x}}(\tau)$)**.** Inspection of the proof shows that

$$\hat{Q}''_{\boldsymbol{x}}(\tau) - Q''_{\boldsymbol{x}}(\tau) = \frac{1}{nh^3}\sum_{i=1}^n \int \boldsymbol{x}^T J(t)^{-1}\boldsymbol{X}_i \left\{t - I\left(U_i \leq t\right)\right\} K''\left(\frac{\tau - t}{h}\right) dt$$

$$+ O_P(n^{-3/4}h^{-2}d^{3/2}\sqrt{\log n}) + O(h^2)$$

uniformly in $(\tau, \boldsymbol{x}) \in [\epsilon, 1-\epsilon]\times\mathcal{X}_0$, and the uniform rate over $(\tau, \boldsymbol{x}) \in [\epsilon, 1-\epsilon]\times\mathcal{X}_0$ of the first term on the right hand side is $O_P(n^{-1/2}h^{-3/2}d\sqrt{\log n})$.

Recall the definition of $\xi_{\boldsymbol{x}}$. In view of the proof of Lemma 8, the following lemma follows relatively directly from Lemma 17.

**Lemma 18.** *Under the conditions of Theorem 4, the following asymptotic linear representation holds uniformly in $\boldsymbol{x} \in \mathcal{X}_0$:*

$$\hat{m}(\boldsymbol{x}) - m(\boldsymbol{x}) = \frac{\sqrt{d}}{nh^{3/2}} \sum_{i=1}^{n} \xi_{\boldsymbol{x}}(U_i, \boldsymbol{X}_i) + O_P(n^{-3/4}h^{-2}d^{3/2}\sqrt{\log n} + n^{-1}h^{-4}d^2 \log n + h^2)$$

*where $U_1, \ldots, U_n \sim U(0,1)$ i.i.d. independent of $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$. In addition, we have*

$$\sup_{\boldsymbol{x} \in \mathcal{X}_0} \left| \frac{1}{nh^{3/2}} \sum_{i=1}^{n} \psi_{\boldsymbol{x}}(U_i, \boldsymbol{X}_i) \right| = O_P(n^{-1/2}h^{-3/2}d\sqrt{\log n}).$$

We are now in position to prove Theorem 4.

*Proof of Theorem 4.* As before, we split the proof into two parts.

**Step 1**. We will apply Proposition 2.1 in Chernozhukov et al. (2017a) to $n^{-1/2}\sum_{i=1}^{n} A\Psi_i$. To this end, we will check Conditions (M.1), (M.2), and (E.1) of Chernozhukov et al. (2017a). Condition (M.1) follows automatically, so we will verify Conditions (M.2) and (E.1).

**Condition (M.2)**. Recall that $\|\boldsymbol{x}\|/\sqrt{d} \leq C_2$ for all $\boldsymbol{x} \in \mathcal{X}_0$. Observe that

$$\begin{aligned}
&\max_{1 \leq k \leq M} \mathbb{E}\left[ \|(\xi_{\boldsymbol{x}_\ell}(U_i, \boldsymbol{X}_i))_{\ell \in S_k}\|_1^3 \right] \\
&\leq O(h^{-3/2}) \sup_{\alpha \in \mathbb{S}^{d-1}} \mathbb{E}[|\alpha^T \boldsymbol{X}|^3] \max_{1 \leq \ell \leq L} \int \left| K''\left( \frac{\tau_{\boldsymbol{x}_\ell} - t}{h} \right) \right|^3 dt = O(h^{-1/2}d^{1/2}),
\end{aligned}$$

where we used the fact that

$$\sup_{\alpha \in \mathbb{S}^{d-1}} \mathbb{E}[|\alpha^T \boldsymbol{X}|^3] \leq C_3\sqrt{d} \sup_{\alpha \in \mathbb{S}^{d-1}} \mathbb{E}[(\alpha^T \boldsymbol{X})^2] = C_3\sqrt{d}\|\mathbb{E}[\boldsymbol{X}\boldsymbol{X}^T]\|_{\text{op}} = O(\sqrt{d}).$$

This implies that $\max_{1 \leq k \leq M} \mathbb{E}[|A_k^T \Psi_i|^3] = O(d^{1/2}h^{-1/2})$. Likewise, $\max_{1 \leq \ell \leq M} \mathbb{E}[|A_k^T \Psi_i|^4] = O(dh^{-1})$.

**Condition (E.2)**. Since $\|\boldsymbol{X}\| \leq C_3\sqrt{d}$, we have $|A_k^T \Psi_i| \leq$ const. $h^{-1/2}d^{1/2}$.

Thus, applying Proposition 2.1 in Chernozhukov et al. (2017a), we have

$$\sup_{b\in\mathbb{R}^M}\left|\mathbb{P}\left(n^{-1/2}\textstyle\sum_{i=1}^n A\Psi_i\le b\right)-\mathbb{P}(AG\le b)\right|\to 0,$$

provided that

$$\frac{d\log^7(Mn)}{nh}\to 0,$$

which is satisfied under our assumption.

**Step 2**. Observe that

$$\begin{aligned}&\left\|A\sqrt{nh^3d^{-1}}(\hat m(\boldsymbol{x}_\ell)-m(\boldsymbol{x}_\ell))_{\ell=1}^L-n^{-1/2}\textstyle\sum_{i=1}^n A\Psi_i\right\|_\infty\\&=O_P(n^{-1/4}h^{-1/2}d\sqrt{\log n}+n^{-1/2}h^{-5/2}d^{3/2}\log n+n^{1/2}h^{7/2}d^{-1/2}).\end{aligned}\tag{B.11}$$

In view of the proof of Step 2 in Theorem 1, the desired conclusion follows if the right hand side on (B.11) is $o_P(1/\sqrt{\log M})$, which is satisfied under our assumption. □

## B.4.2 Proof of Theorem 5

Define $\check\Psi_i:=(\check\psi_{\boldsymbol{x}_1}(U_i,\boldsymbol{X}_i),\dots,\check\psi_{\boldsymbol{x}_L}(U_i,\boldsymbol{X}_i))^T$ with

$$\check\psi_{\boldsymbol{x}}(u,\boldsymbol{x}'):=\frac{s_{\boldsymbol{x}}(\tau_{\boldsymbol{x}})}{s''_{\boldsymbol{x}}(\tau_{\boldsymbol{x}})\sqrt{dh}}K'\left(\frac{\tau_{\boldsymbol{x}}-u}{h}\right)\boldsymbol{x}^TJ(\tau_{\boldsymbol{x}})^{-1}\boldsymbol{x}'.$$

Further, define $\check\Sigma=\mathbb{E}\left[\check\Psi_i\check\Psi_i^T\right]$, $\check\Gamma:=\mathrm{diag}\{\check\sigma_1,\dots,\check\sigma_M\}$ with $\check\sigma_i^2:=D_i^T\check\Sigma D_i$, and $\check A:=\check\Gamma^{-1}D$.

The following operator norm bound is in parallel to Lemma 10 for the fixed dimensional case.

**Lemma 19.** *Under the conditions of Theorem 5, we have*

$$\sup_{\boldsymbol{x}\in\mathcal{X}_0}\|\hat J(\hat\tau_{\boldsymbol{x}})-J(\tau_{\boldsymbol{x}})\|_{\mathrm{op}}=O_P(n^{-1/2}h^{-3/2}d^{3/2}\sqrt{\log n}+h^2).$$

*Proof.* Observe that the left hand side can be bounded by

$$\begin{aligned}
&\left\| \frac{1}{n}\sum_{i=1}^{n} K_h(Y_i - \boldsymbol{X}_i^T\hat{\beta}(\hat{\tau}_{\boldsymbol{x}}))(\alpha^T\boldsymbol{X}_i)^2 - \mathbb{E}\left[K_h(Y_i - \boldsymbol{X}_i^T\beta)(\alpha^T\boldsymbol{X})^2\right]\big|_{\beta=\hat{\beta}(\hat{\tau}_{\boldsymbol{x}})} \right\|_{\mathbb{S}^{d-1}\times\mathcal{X}_0} \\
&\quad + \left\| \mathbb{E}\left[K_h(Y_i - \boldsymbol{X}_i^T\beta)(\alpha^T\boldsymbol{X})^2\right]\big|_{\beta=\hat{\beta}(\hat{\tau}_{\boldsymbol{x}})} - \mathbb{E}\left[f(\boldsymbol{X}^T\beta \mid \boldsymbol{X})(\alpha^T\boldsymbol{X})^2\right]\big|_{\beta=\hat{\beta}(\hat{\tau}_{\boldsymbol{x}})} \right\|_{\mathbb{S}^{d-1}\times\mathcal{X}_0} \\
&\quad + \left\| \mathbb{E}\left[f(\boldsymbol{X}^T\beta \mid \boldsymbol{X})(\alpha^T\boldsymbol{X})^2\right]\big|_{\beta=\hat{\beta}(\hat{\tau}_{\boldsymbol{x}})} - \mathbb{E}\left[f(\boldsymbol{X}^T\beta(\tau_{\boldsymbol{x}}) \mid \boldsymbol{X})(\alpha^T\boldsymbol{X})^2\right] \right\|_{\mathbb{S}^{d-1}\times\mathcal{X}_0} \\
&=: I + II + III,
\end{aligned}$$

where $\|\cdot\|_{\mathbb{S}^{d-1}\times\mathcal{X}_0} = \sup_{(\alpha,\beta)\in\mathbb{S}^{d-1}\times\mathcal{X}_0}|\cdot|$. By Taylor expansion and $\sup_{\alpha\in\mathbb{S}^{d-1}}\mathbb{E}[(\alpha^T\boldsymbol{X})^2] = \|\mathbb{E}[\boldsymbol{X}\boldsymbol{X}^T]\|_{\mathrm{op}} = O(1)$, we see that $II = O(h^2)$. Next, applying the local maximal inequality (Lemma 2) combined with the fact that $\sup_{\alpha\in\mathbb{S}^{d-1}}\mathbb{E}[|\alpha^T\boldsymbol{X}|^4] = O(d)$, we can show that $I = O_P(\sqrt{n^{-1}h^{-1}d^2\log n})$. Finally, the term $III$ is bounded by

$$C_1\|\hat{\beta}(\hat{\tau}_{\boldsymbol{x}}) - \beta(\tau_{\boldsymbol{x}})\|_{\mathcal{X}_0} \underbrace{\sup_{\alpha\in\mathbb{S}^{d-1}}\mathbb{E}[|\alpha^T\boldsymbol{X}|^3]}_{=O(d^{1/2})} \quad \text{and}$$

$$\begin{aligned}
\|\hat{\beta}(\hat{\tau}_{\boldsymbol{x}}) - \beta(\tau_{\boldsymbol{x}})\|_{\mathcal{X}_0} &= O_P\left(\|\hat{\beta} - \beta\|_{[\epsilon,1-\epsilon]} \bigvee \|\beta(\hat{\tau}_{\boldsymbol{x}}) - \beta(\tau_{\boldsymbol{x}})\|_{\mathcal{X}_0}\right) \\
&= O_P\left(n^{-1/2}d^{1/2} \bigvee n^{-1/2}h^{-3/2}d\sqrt{\log n}\right) = O_P(n^{-1/2}h^{-3/2}d\sqrt{\log n}),
\end{aligned}$$

where we used the observation that $Q'_{\boldsymbol{x}}(\tau) = \boldsymbol{x}^T\beta'(\tau)$ is bounded in $(\tau,\boldsymbol{x}) \in [\epsilon, 1-\epsilon]\times\mathcal{X}$. Conclude that $III = O_P(n^{-1/2}h^{-3/2}d^{3/2}\sqrt{\log n})$. □

Similarly, we have the following lemma in parallel to Lemma 11.

**Lemma 20.** *Under Assumption 3, we have*

$$\|\hat{\Sigma} - \check{\Sigma}\|_\infty = O_P(n^{-1/2}h^{-3/2}d(d^{1/2}\vee h^{-1})\sqrt{\log n} + h).$$

*Proof.* The proof is analogous to the proof of Lemma 11, given that $\boldsymbol{x}_\ell/\sqrt{d} \le C_2$ and we added normalization by $\sqrt{d}$ in the definition of $\psi_{\boldsymbol{x}}$. The only missing

part is a bound on

$$\left\| \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i^T - \mathbb{E}[\boldsymbol{X}\boldsymbol{X}^T] \right\|_{\text{op}},$$

but Rudelson's inequality yields that the above term is $O_P(\sqrt{d(\log d)/n})$; cf. Rudelson (1999).

□

We are now in position to prove Theorem 5.

*Proof of Theorem 5.* Observe that

$$\begin{aligned} &\sup_{b \in \mathbb{R}^M} \left| \mathbb{P}_{|\mathcal{D}_n} \left( n^{-1/2} \textstyle\sum_{i=1}^n \hat{A}\hat{\Psi}_i \leq b \right) - \mathbb{P}(AG \leq b) \right| \\ &\leq \sup_{b \in \mathbb{R}^M} \left| \mathbb{P}_{|\mathcal{D}_n} \left( n^{-1/2} \textstyle\sum_{i=1}^n \hat{A}\hat{\Psi}_i \leq b \right) - \mathbb{P}(\check{A}\check{G} \leq b) \right| \\ &\quad + \sup_{b \in \mathbb{R}^M} \left| \mathbb{P}(\check{A}\check{G} \leq b) - \mathbb{P}(AG \leq b) \right|, \end{aligned} \tag{B.12}$$

where $\check{G} \sim N(0, \check{\Sigma})$. The first term on the right hand side of (B.12) is bounded by

$$\underbrace{\sup_{b \in \mathbb{R}^M} \left| \mathbb{P}_U \left( n^{-1/2} \textstyle\sum_{i=1}^n \hat{A}\hat{\Psi}_i \leq b \right) - \mathbb{P}_U(\hat{A}\hat{G} \leq b) \right|}_{I}$$

$$+ \underbrace{\sup_{b \in \mathbb{R}^M} \left| \mathbb{P}_U(\hat{A}\hat{G} \leq b) - \mathbb{P}_{|\mathcal{D}_n}(\check{A}\hat{G} \leq b) \right|}_{II} + \underbrace{\sup_{b \in \mathbb{R}^M} \left| \mathbb{P}_U(\check{A}\hat{G} \leq b) - \mathbb{P}(\check{A}\check{G} \leq b) \right|}_{III},$$

where $\hat{G} \sim N(0, \hat{\Sigma})$ conditionally on $\mathcal{D}_n$. For $I$, we can apply Proposition 2.1 in Chernozhukov et al. (2017a) conditionally on $\mathcal{D}_n$. Similarly to the last part of the proof of Theorem 2, we can show that $I = o_P(1)$ if

$$\frac{d \log^7 (Mn)}{nh} \to 0,$$

which is satisfied under our assumption. We can analyze $II$ and $III$ as in the proof of Theorem 2 and show that $II \vee III = o_P(1)$ if $n^{-1/2} h^{-3/2} d(d^{1/2} \vee$

$h^{-1})(\sqrt{\log n})\log^2 M = o(1)$ and $h\log^2 M = o(1)$, which is satisfied under our assumption.

Finally, in view of the Gaussian comparison inequality (Lemma 4), we see that the second term on the right hand side of (B.12) is $o(1)$ if $\|\check{A}\check{\Sigma}\check{A}^T - A\Sigma A^T\|_\infty \log^2 M = o(1)$. It is not difficult to see that

$$\|\check{A}\check{\Sigma}\check{A}^T - A\Sigma A^T\|_\infty = O\left(\sup_{\boldsymbol{x}_1,\boldsymbol{x}_2\in\mathcal{X}_0} |\mathbb{E}[\xi_{\boldsymbol{x}_1}\xi_{\boldsymbol{x}_2} - \check{\psi}_{\boldsymbol{x}_1}\check{\psi}_{\boldsymbol{x}_2}]|\right) = O(h) = o(1/\log^2 M).$$

This completes the proof. □

## B.5 Additional simulation results

### B.5.1 Nonparametric bootstrap pointwise confidence intervals

In this section, we present simulation results for the nonparametric bootstrap. Due to the heavy computational burden of the nonparametric bootstrap, we only consider pointwise confidence intervals in the simulation. We consider the $lmNormal$, $lmLognormal$ and $Nonlinear$ models as in Section 2.4.1, together with the same subsample sizes, $n = 500$, $1000$ and $2000$, and repetition number $s = 500$. The results are presented in Tables B.1–B.3.

From the simulation results, the nonparametric bootstrap confidence intervals achieve close to nominal coverage probabilities under large sample sizes for the *lmNormal* and *Nonlinear* models. For the *lmLognormal* model, the nonparametric bootstrap confidence intervals have lower coverage probabilities than the nominal level. This may be due to the slow convergence rate of the bootstrap

Table B.1: Nonparametric bootstrap pointwise confidence intervals for *lmNormal* model.

| Design point | Sample size | Coverage probability | | Median length | | IQR | |
|---|---|---|---|---|---|---|---|
| | | 95% | 99% | 95% | 99% | 95% | 99% |
| $X_1$=0.3 | $n = 500$ | 92.4% | 97.6% | 0.80 | 1.14 | 0.38 | 0.52 |
| | $n = 1000$ | 93.2% | 98.8% | 0.69 | 0.95 | 0.29 | 0.39 |
| | $n = 2000$ | 92% | 99.2% | 0.57 | 0.79 | 0.25 | 0.33 |
| $X_1$=0.5 | $n = 500$ | 92.4% | 98.4% | 0.91 | 1.25 | 0.40 | 0.57 |
| | $n = 1000$ | 93.4% | 98.2% | 0.76 | 1.05 | 0.30 | 0.44 |
| | $n = 2000$ | 93.4% | 98% | 0.64 | 0.88 | 0.22 | 0.32 |
| $X_1$=0.7 | $n = 500$ | 92.4% | 98.6% | 1.10 | 1.57 | 0.54 | 0.80 |
| | $n = 1000$ | 91.4% | 97.6% | 0.91 | 1.26 | 0.40 | 0.58 |
| | $n = 2000$ | 93.8% | 97.6% | 0.79 | 1.07 | 0.27 | 0.40 |

Table B.2: Nonparametric bootstrap pointwise confidence intervals for *lmLognormal* model.

| Design point | Sample size | Coverage probability | | Median length | | IQR | |
|---|---|---|---|---|---|---|---|
| | | 95% | 99% | 95% | 99% | 95% | 99% |
| $X_1$=0.3 | $n = 500$ | 87.8% | 95.8% | 2.85 | 4.31 | 2.70 | 4.06 |
| | $n = 1000$ | 86.2% | 94.4% | 1.86 | 3.27 | 2.05 | 3.61 |
| | $n = 2000$ | 89.2% | 95.4% | 1.47 | 2.69 | 2.07 | 3.36 |
| $X_1$=0.5 | $n = 500$ | 85% | 95.2% | 3.28 | 5.05 | 2.86 | 4.14 |
| | $n = 1000$ | 88% | 95.8% | 2.03 | 3.57 | 2.39 | 3.99 |
| | $n = 2000$ | 84.4% | 92.6% | 1.55 | 2.76 | 1.99 | 3.49 |
| $X_1$=0.7 | $n = 500$ | 82% | 92.8% | 4.51 | 6.64 | 4.27 | 6.69 |
| | $n = 1000$ | 86.2% | 96% | 2.76 | 5.39 | 3.10 | 4.43 |
| | $n = 2000$ | 83.6% | 94.4% | 1.75 | 3.13 | 2.07 | 3.95 |

approximation to the sampling distribution under such a data generating process. Compared with the pivotal bootstrap confidence intervals in the other two models, the nonparametric bootstrap provides shorter and more stable confidence intervals, i.e., less variable interval lengths, in the *lmNormal* model while the pivotal bootstrap performs better in the *Nonlinear* model.

To further demonstrate the computational advantage of the pivotal boot-

Table B.3: Nonparametric bootstrap pointwise confidence intervals for *Nonlinear* model.

| Design point | Sample size | Coverage probability | | Median length | | IQR | |
|---|---|---|---|---|---|---|---|
| | | 95% | 99% | 95% | 99% | 95% | 99% |
| $X_1$=0.7 | $n = 500$ | 94.8% | 99% | 1.05 | 1.62 | 1.55 | 2.62 |
| | $n = 1000$ | 92.2% | 98.4% | 0.78 | 1.27 | 1.01 | 1.91 |
| | $n = 2000$ | 94.8% | 98.8% | 0.67 | 1.08 | 0.92 | 1.32 |
| $X_1$=0.9 | $n = 500$ | 94% | 99.4% | 1.11 | 1.59 | 1.25 | 1.77 |
| | $n = 1000$ | 94.0% | 98% | 0.84 | 1.42 | 1.18 | 1.57 |
| | $n = 2000$ | 94.8% | 98% | 0.60 | 1.04 | 0.86 | 1.25 |
| $X_1$=1.1 | $n = 500$ | 94% | 99% | 0.80 | 1.49 | 1.05 | 1.17 |
| | $n = 1000$ | 95.2% | 99.2% | 0.56 | 1.08 | 0.80 | 1.14 |
| | $n = 2000$ | 92.8% | 97.8% | 0.36 | 0.60 | 0.39 | 0.74 |

strap over the nonparametric bootstrap, we report the average running time of these two bootstraps in the *Nonlinear* model with design point $X_1 = 0.9$ (results in other scenarios are similar). The simulation results are obtained in the R environment with 28 Intel Xeon processors and 240 Gbytes RAM over Red Hat OpenStack Platform. We measure the average running time in seconds and report the results in Table B.4. From the table, we can see that the pivotal bootstrap requires substantially less computational time than the nonparametric bootstrap as predicted in Remark 6 in the main text.

Table B.4: Running time comparison between the pivotal and nonparametric bootstraps.

| Design point | Sample size | Pivotal bootstrap | Nonparametric bootstrap |
|---|---|---|---|
| $X_1$=0.9 | $n = 500$ | 1.02 | 22.23 |
| | $n = 1000$ | 1.43 | 33.79 |
| | $n = 2000$ | 2.27 | 58.08 |

### B.5.2 Mean squared error comparison with existing modal estimators

In the following, we present the mean squared error of our modal estimator $\hat{m}_{\boldsymbol{x}}$, which is defined by

$$MSE(\hat{m}_{\boldsymbol{x}}) := s^{-1} \sum_{i=1}^{s} (\hat{m}_{\boldsymbol{x}}^{(i)} - m_{\boldsymbol{x}})^2,$$

where $m_{\boldsymbol{x}}$ is the true conditional mode and $\hat{m}_{\boldsymbol{x}}^{(i)}$ is our modal estimator in the $i$-th repetition. We compare our method with existing methods by Ota et al. (2019), Kemp and Santos-Silva (2012) and Yao and Li (2014). We consider the $lmNormal$, $lmLognormal$ and $Nonlinear$ models as in Section 2.4.1. The same subsample sizes, $n = 500$, $1000$ and $2000$, and repetition number $s = 500$ are considered. The results are presented in Tables B.5–B.7. The KS-YL estimator refers to the linear modal estimator studied by Kemp and Santos-Silva (2012) and Yao and Li (2014).

Table B.5: Mean squared error comparison: *lmNormal*.

| Design point | Sample size | Our estimator | Ota et al. (2019) | KS-YL estimator |
|---|---|---|---|---|
| $X_1$=0.3 | $n = 500$ | 0.020 | 0.063 | 0.207 |
| | $n = 1000$ | 0.018 | 0.043 | 0.208 |
| | $n = 2000$ | 0.010 | 0.032 | 0.275 |
| $X_1$=0.5 | $n = 500$ | 0.026 | 0.091 | 0.712 |
| | $n = 1000$ | 0.017 | 0.063 | 0.713 |
| | $n = 2000$ | 0.012 | 0.047 | 0.928 |
| $X_1$=0.7 | $n = 500$ | 0.109 | 0.167 | 1.559 |
| | $n = 1000$ | 0.026 | 0.128 | 1.840 |
| | $n = 2000$ | 0.024 | 0.103 | 1.954 |

From the simulation results, while no method dominates in all the three scenarios, the proposed modal estimator performs reasonably well uniformly in all

Table B.6: Mean squared error comparison: *lmLognormal*.

| Design point | Sample size | Our estimator | Ota et al. (2019) | KS-YL estimator |
|---|---|---|---|---|
| $X_1$=0.3 | $n = 500$ | 0.286 | 0.329 | 0.146 |
| | $n = 1000$ | 0.162 | 0.296 | 0.124 |
| | $n = 2000$ | 0.118 | 0.259 | 0.066 |
| $X_1$=0.5 | $n = 500$ | 0.421 | 0.508 | 0.204 |
| | $n = 1000$ | 0.219 | 0.466 | 0.160 |
| | $n = 2000$ | 0.151 | 0.410 | 0.094 |
| $X_1$=0.7 | $n = 500$ | 0.845 | 0.774 | 0.407 |
| | $n = 1000$ | 0.491 | 0.657 | 0.269 |
| | $n = 2000$ | 0.217 | 0.605 | 0.224 |

Table B.7: Mean squared error comparison: *Nonlinear*.

| Design point | Sample size | Our estimator | Ota et al. (2019) | KS-YL estimator |
|---|---|---|---|---|
| $X_1$=0.3 | $n = 500$ | 0.0090 | 0.0023 | 0.051 |
| | $n = 1000$ | 0.0071 | 0.00082 | 0.041 |
| | $n = 2000$ | 0.0044 | 0.00045 | 0.031 |
| $X_1$=0.5 | $n = 500$ | 0.016 | 0.0052 | 0.075 |
| | $n = 1000$ | 0.010 | 0.0046 | 0.063 |
| | $n = 2000$ | 0.0086 | 0.0043 | 0.049 |
| $X_1$=0.7 | $n = 500$ | 0.026 | 0.025 | 0.074 |
| | $n = 1000$ | 0.017 | 0.023 | 0.068 |
| | $n = 2000$ | 0.012 | 0.022 | 0.052 |

the settings. In particular, our estimator and Ota et al. (2019)'s estimator outperform the KS-YL linear modal estimator in the *Nonlinear* model as expected.

### B.5.3 Simulation results for the pivotal bootstrap testing

In this section, we consider testing significance of a covariate on the conditional mode. Suppose covariate $\boldsymbol{X} = (1, X_1, X_2)$ where $X_1$ is continuous and $X_2$ is binary (0 or 1). We want to test the null hypothesis $H_0 : m(X_1, 0) = m(X_1, 1)$ versus the alternative hypothesis $H_1 : m(X_1, 0) \neq m(X_1, 1)$, where $m(x_1, x_2)$ is

the conditional mode of $Y$ given $X_1 = x_1$ and $X_2 = x_2$. We will generate $\boldsymbol{X}$ according to $X_1 \sim \text{Unif}(0, 1)$ and $X_2 \sim \text{Binomial}(0.5)$. For the outcome $Y$, two generation schemes are considered: (1) $Y = 1+3X_1+\xi$ and (2) $Y = 1+3X_1+\alpha X_2+\xi$, where we take $\xi \sim N(0, 1)$ and $\alpha \neq 0$ in both models. The corresponding mode functions are $m(\boldsymbol{X}) = 1 + 3X_1$ and $m(\boldsymbol{X}) = 1 + 3X_1 + \alpha X_2$, respectively. Therefore, the two generation schemes correspond to $H_0$ being true and false respectively which allows us to evaluate both power and size of our bootstrap testing procedure. We will take $\alpha = 0.8$ or $1$ in the simulation. In the current setup, the limiting Gaussian distribution given by Theorem 1 is one dimensional and the corresponding variance can be calculated explicitly based on the above setup. Therefore, an oracle test procedure can be constructed by using the quantiles of the corresponding limiting Gaussian distribution to define the test rejection region. We will compare the performance of our bootstrap testing with this benchmark oracle testing.

We conduct hypothesis testing of nominal level $0.05$ and $0.01$ for $X_1$ taking value at 0.3, 0.5 and 0.7. For each value of $X_1$, three different sample sizes from 500 to 2000 are considered. We report the empirical size and power of both bootstrap testing and oracle testing based on 500 simulations in Tables B.8 and B.9.

From the tables, the Type I errors are well preserved for both tests at three design points while the bootstrap testing committed slightly fewer Type I errors. We can see the decrease of power of both bootstrap testing and oracle testing under the same design point ($X_1$) and subsample size ($n$) as $\alpha$ gets smaller. For a fixed design point, the power of both tests approaches $1$ with the increasing subsample size which supports our theory. In particular, the good performance

Table B.8: Size and power for bootstrap testing and oracle testing ($\alpha = 1$).

| Design Point | Sample size | Bootstrap testing | | | | Oracle testing | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Size | | Power | | Size | | Power | |
| | | 0.05 | 0.01 | 0.05 | 0.01 | 0.05 | 0.01 | 0.05 | 0.01 |
| $X_1 = 0.3$ | $n = 500$ | 0 | 0 | 0.622 | 0.348 | 0.012 | 0 | 1 | 0.996 |
| | $n = 1000$ | 0.006 | 0.002 | 0.814 | 0.628 | 0.022 | 0.004 | 1 | 1 |
| | $n = 2000$ | 0 | 0 | 0.932 | 0.874 | 0.032 | 0.002 | 1 | 1 |
| $X_1 = 0.5$ | $n = 500$ | 0.004 | 0 | 0.55 | 0.322 | 0.016 | 0.002 | 0.994 | 0.984 |
| | $n = 1000$ | 0.004 | 0 | 0.794 | 0.608 | 0.048 | 0.002 | 1 | 0.996 |
| | $n = 2000$ | 0.004 | 0 | 0.928 | 0.84 | 0.022 | 0.006 | 1 | 1 |
| $X_1 = 0.7$ | $n = 500$ | 0.002 | 0 | 0.596 | 0.408 | 0.012 | 0.02 | 0.998 | 0.994 |
| | $n = 1000$ | 0 | 0 | 0.816 | 0.644 | 0.024 | 0.006 | 0.998 | 0.996 |
| | $n = 2000$ | 0.004 | 0 | 0.928 | 0.842 | 0.038 | 0.006 | 1 | 1 |

Table B.9: Power for bootstrap testing and oracle testing with $\alpha = 0.8$.

| Design point | Sample size | Bootstrap testing | | Oracle testing | |
|---|---|---|---|---|---|
| | | 0.05 | 0.01 | 0.05 | 0.01 |
| $X_1 = 0.3$ | $n = 500$ | 0.39 | 0.176 | 0.978 | 0.894 |
| | $n = 1000$ | 0.66 | 0.406 | 0.992 | 0.98 |
| | $n = 2000$ | 0.858 | 0.696 | 1 | 1 |
| $X_1 = 0.5$ | $n = 500$ | 0.392 | 0.18 | 0.962 | 0.846 |
| | $n = 1000$ | 0.628 | 0.396 | 0.992 | 0.968 |
| | $n = 2000$ | 0.874 | 0.706 | 1 | 0.998 |
| $X_1 = 0.7$ | $n = 500$ | 0.416 | 0.21 | 0.976 | 0.894 |
| | $n = 1000$ | 0.66 | 0.432 | 0.996 | 0.982 |
| | $n = 2000$ | 0.872 | 0.708 | 1 | 0.998 |

of the oracle testing justifies our normal approximation theory. The performance of the proposed bootstrap testing is inferior to the oracle testing under the same design point and subsample size. This may due to several reasons including the bootstrap approximation error and the bias in the estimation of the nuisance parameters. However, the performance of bootstrap testing is reasonable when the sample size is sufficiently large which agrees with our asymptotic theory.

We also remark that we here test the significance of the covariates for the con-

ditional mode by testing the change of the conditional mode due to the change of design points, instead of testing the corresponding coefficient in the quantile regression slope vector. This is due to the following observation. Although we assume a linear quantile model, it does not imply a direct modeling on the conditional mode function. That is, the conditional mode depends on the covariates in an implicit way under our quantile-based mode regression, i.e. under our modeling, the mode function is $m(\boldsymbol{x}) = \boldsymbol{x}^T\beta(\tau_{\boldsymbol{x}})$ where the coordinates of $\beta(\tau_{\boldsymbol{x}})$ are functions of $\boldsymbol{x}$ that may have arbitrary forms (as long as they satisfy some natural restrictions resulted from quantile functions). It is possible that one coefficient of the $\beta(\tau_{\boldsymbol{x}})$ is nonzero while the corresponding covariate does not contribute to the conditional mode function. This is in contrast to the linear modal regression where the mode function is assumed to be $m(\boldsymbol{x}) = \boldsymbol{x}^T\beta$ where $\beta$ is a constant vector (independent of $\boldsymbol{x}$) and therefore testing the significance of the covariates is equivalent to testing the coefficients in $\beta$ being $0$ or not.

### B.5.4 Pivotal bootstrap confidence intervals using oracle information

In this section, we provide simulation results of pivotal bootstrap inference using oracle model information, i.e., we estimate the nuisance parameters in the influence function based on the underlying true density or conditional quantile function. We reexamine the setups in Section 2.4.1 and the corresponding results for pointwise confidence intervals are presented in Tables B.10 to B.12 and Table B.13 for approximate confidence bands.

From the simulation results, the oracle confidence intervals achieve close to

Table B.10: Oracle pointwise confidence intervals for *lmNormal* model.

| Design point | Sample size | Coverage probability | | Median length | | IQR | |
|---|---|---|---|---|---|---|---|
| | | 95% | 99% | 95% | 99% | 95% | 99% |
| $X_1$=0.3 | $n = 500$ | 95.4% | 99.6% | 0.94 | 1.24 | 0.16 | 0.19 |
| | $n = 1000$ | 95% | 99.2% | 0.79 | 1.02 | 0.13 | 0.14 |
| | $n = 2000$ | 96.6% | 99% | 0.64 | 0.85 | 0.08 | 0.10 |
| $X_1$=0.5 | $n = 500$ | 95.2% | 99.2% | 1.04 | 1.38 | 0.24 | 0.32 |
| | $n = 1000$ | 94.6% | 98.6% | 0.86 | 1.14 | 0.15 | 0.20 |
| | $n = 2000$ | 95% | 99.2% | 0.71 | 0.94 | 0.11 | 0.13 |
| $X_1$=0.7 | $n = 500$ | 96.6% | 98.8% | 1.29 | 1.70 | 0.37 | 0.41 |
| | $n = 1000$ | 96% | 99.4% | 1.06 | 1.42 | 0.22 | 0.27 |
| | $n = 2000$ | 95.6% | 99.6% | 0.88 | 1.16 | 0.16 | 0.21 |

Table B.11: Oracle pointwise confidence intervals for *lmLognormal* model.

| Design point | Sample size | Coverage probability | | Median length | | IQR | |
|---|---|---|---|---|---|---|---|
| | | 95% | 99% | 95% | 99% | 95% | 99% |
| $X_1$=0.3 | $n = 500$ | 77% | 80.6% | 1.44 | 1.86 | 0.91 | 1.20 |
| | $n = 1000$ | 94% | 96% | 1.34 | 1.75 | 0.52 | 0.63 |
| | $n = 2000$ | 96.4% | 99.8% | 1.01 | 1.33 | 0.28 | 0.35 |
| $X_1$=0.5 | $n = 500$ | 83.4% | 84.8% | 2.07 | 2.67 | 1.31 | 1.77 |
| | $n = 1000$ | 96% | 97.8% | 1.85 | 2.43 | 0.79 | 1.00 |
| | $n = 2000$ | 97.4% | 99.8% | 1.52 | 2.02 | 0.54 | 0.69 |
| $X_1$=0.7 | $n = 500$ | 76.6% | 81% | 2.33 | 3.19 | 1.65 | 2.26 |
| | $n = 1000$ | 93.4% | 94.8% | 2.25 | 3.03 | 1.36 | 1.61 |
| | $n = 2000$ | 96.2% | 98.6% | 1.84 | 2.43 | 0.76 | 0.92 |

nominal level coverage probabilities when the sample size is sufficiently large, which supports our theoretical results.

Table B.12: Oracle pointwise confidence intervals for *Nonlinear* model.

| Design point | Sample size | Coverage probability | | Median length | | IQR | |
|---|---|---|---|---|---|---|---|
| | | 95% | 99% | 95% | 99% | 95% | 99% |
| $X_1$=0.7 | $n = 500$ | 92.4% | 97.4% | 0.58 | 0.73 | 0.48 | 0.47 |
| | $n = 1000$ | 96% | 98.8% | 0.45 | 0.59 | 0.45 | 0.49 |
| | $n = 2000$ | 97% | 99% | 0.35 | 0.46 | 0.22 | 0.29 |
| $X_1$=0.9 | $n = 500$ | 94.4% | 97.2% | 0.62 | 0.81 | 0.46 | 0.57 |
| | $n = 1000$ | 94.8% | 97.6% | 0.50 | 0.66 | 0.24 | 0.32 |
| | $n = 2000$ | 96% | 98.8% | 0.45 | 0.58 | 0.19 | 0.24 |
| $X_1$=1.1 | $n = 500$ | 93.2% | 96.8% | 0.53 | 0.70 | 0.18 | 0.26 |
| | $n = 1000$ | 94.8% | 97.6% | 0.46 | 0.59 | 0.17 | 0.21 |
| | $n = 2000$ | 94% | 98% | 0.37 | 0.48 | 0.10 | 0.14 |

Table B.13: Oracle approximate confidence bands for *lmNormal*, *lmLognormal* and *Nonlinear* models.

| Models | Sample size | Coverage probability | | Median length | |
|---|---|---|---|---|---|
| | | 95% | 99% | 95% | 99% |
| lmNormal | $n = 500$ | 94.8% | 98.6% | 1.14 | 1.48 |
| | $n = 1000$ | 94.6% | 98.8% | 0.97 | 1.22 |
| | $n = 2000$ | 93.4% | 98.8% | 0.79 | 1.02 |
| lmLognormal | $n = 500$ | 82.6% | 84.2% | 2.39 | 3.06 |
| | $n = 1000$ | 95.2% | 97.1% | 2.03 | 2.65 |
| | $n = 2000$ | 97.6% | 99.4% | 1.73 | 2.18 |
| Nonlinear | $n = 500$ | 94.6% | 97.6% | 1.12 | 1.23 |
| | $n = 1000$ | 96% | 97.8% | 0.82 | 0.98 |
| | $n = 2000$ | 94.2% | 98% | 0.49 | 0.59 |

## B.6 Additional discussion: model misspecification and quantile crossing

In theory, the quantile crossing problem does not happen since we assume that

$$s_{\boldsymbol{x}}(\tau) = Q'_{\boldsymbol{x}}(\tau) = 1/f(Q_{\boldsymbol{x}}(\tau) \mid \boldsymbol{x}) \geq 1/c_1 > 0,\ \tau \in [\epsilon, 1-\epsilon].$$

See Condition (v) in Assumption 1. This implies the quantile function is strictly increasing in the quantile index $\tau \in [\epsilon, 1-\epsilon]$. Further, under our assumption,

$\hat{s}_{\boldsymbol{x}}(\tau) = \hat{Q}'_{\boldsymbol{x}}(\tau)$ is uniformly consistent over $\tau \in [\epsilon, 1-\epsilon]$ and $\boldsymbol{x} \in \mathcal{X}_0$ (see Lemma 7 in Appendix), so that the estimated conditional quantile function $\tau \mapsto \hat{Q}_{\boldsymbol{x}}(\tau)$ is strictly increasing on $[\epsilon, 1-\epsilon]$ with probability approaching one.

Of course, the quantile crossing problem may happen in the finite sample even if the model is correctly specified (Koenker, 2005). Several methods have been proposed to deal with the quantile crossing problem in the literature, cf. He (1997), Chernozhukov et al. (2010) and Bondell et al. (2010). We can apply any of such monotonization methods to the estimated conditional quantile function to prevent quantile crossing; our theoretical results continue to hold for such a monotonized conditional quantile estimate, as the monotonized estimate agrees with the vanilla conditional quantile estimate with probability approaching one.

Under model misspecification, the fitted linear quantile function can be interpreted as the best linear approximation to the quantile function under some quadratic discrepancy measure cf. section 2.9 in Koenker (2005) and Angrist et al. (2006). Thus, our $\tau_x$ can be interpreted as a minimization point of the derivative of the best linear approximation. Significant number of quantile crossings also implies the misspecification of the model (Koenker, 2005). If this happens, we suggest using the series approximation approach of our method as discussed in Section 5.

## B.7 More implementation details of Section 2.4.1

In this section, we provide more implementation details of estimating $f^{(2)}(m(\boldsymbol{x}) \mid \boldsymbol{x})$ in the simulation. We use the following kernel estimator for

$f^{(2)}(m(\boldsymbol{x}) \mid \boldsymbol{x})$ (recall that $\boldsymbol{X} = (1, X_1)$):

$$\hat{f}^{(2)}(\hat{m}(\boldsymbol{x}) \mid \boldsymbol{x}) = \frac{(nb_Y^3 b_{X_1})^{-1} \sum_{i=1}^n K_1''((\hat{m}(\boldsymbol{x}) - Y_i)/b_Y) K_2((x_1 - X_{i1})/b_{X_1})}{(nb_{X_1})^{-1} \sum_{i=1}^n K_2((x_1 - X_{i1})/b_{X_1})},$$

where $X_{i1}$ is the observed value of $X_1$ in the $i$-th data point. We use the Gaussian kernel for $K_1$ and the Epanechnikov kernel for $K_2$ in the simulation. We use the following bandwidths for covariate $X_1$ and $Y$ in our simulation, respectively:

$$b_{X_1} = \omega \cdot n^{-1/5} \hat{\sigma}_X \quad \text{and} \quad b_Y = \omega \cdot n^{-1/9} \hat{\sigma}_Y, \tag{B.13}$$

where $\hat{\sigma}_\cdot$ is the corresponding sample standard deviation and $\omega > 0$. To select $\omega$ in a data-driven approach, we propose the following procedure. The procedure is motivated by the $L_\infty$-based bandwidth selector in Bissantz et al. (2007).

**Step 1**. Choose a proper grid $G_1$ with $J$ values for $\omega$.

**Step 2**. Generate a dataset $\mathcal{D}_n$ of size $n$ and then compute $\hat{f}^{(2)}(\hat{m}(\boldsymbol{x}) \mid \boldsymbol{x})$ by taking $\omega = G_1(j)$ $(1 \le j \le J)$ and denote the resulting estimator by $\hat{f}_j^{(2)}$. Compute $d_{j,j+1} := \left|\hat{f}_{j+1}^{(2)} - \hat{f}_j^{(2)}\right|$ $(1 \le j \le J-1)$. Choose $\omega^* := \max\{G_1(j) : d_{j,j+1} \ge t \cdot d_{J-1,J}, 1 \le j \le J-1\}$ for some $t > 1$. Repeat the computation of $\omega^*$ for $N$ times to get $\omega_1^*, \cdots, \omega_N^*$ and denote the mode of $\{\omega_1^*, \cdots, \omega_N^*\}$s by $\omega_{opt}^*$.

**Step 3**. Choose a subgrid $G_2 \subset G_1$ centering at $\omega_{opt}^*$. Then we proceed as Step 2 and output the final selected $\omega$.

In fact, we may iterate the above procedure for more times by using a further subgrid based on the output of Step 3. However, we find the above three-step algorithm works reasonably well in our numerical experiments. In our simulation of pointwise confidence intervals, we take $G_1$ as a equally spaced grid on $[0.05, 1.25]$ with $J = 13$; $2 \le t \le 4$ ($t = 2$ for *lmNormal* and $3 \le t \le 4$ for *lmLogNormal* and *Nonlinear*); $N = 500$ and $G_2$ of length 7. To relieve the computational

burden, for each model we only select $\omega$ once for a particular design point $\boldsymbol{x}$ at sample size $n = 2000$. The selected values of $\omega$ for pointwise confidence intervals are reported in Tables B.14 and B.15. For the simulation of confidence bands, we use a common $\omega$ for all the different design points in the considered models to reduce the computational burden and the corresponding values are reported in Table B.16. Based on our numerical experience, we recommend a slightly larger $\omega$ than the pointwisely selected $\omega$ for the confidence bands if a common $\omega$ is used. In practice, we may use pointwisely selected $\omega$ for different design points when constructing confidence bands. For the practical use, $\mathcal{D}_n$ in Step 2 may be taken as the bootstrap subsamples.

Table B.14: Values of $\omega$ selected for *lmNormal* and *lmLognormal* models

| Models | $X_1 = 0.3$ | $X_1 = 0.5$ | $X_1 = 0.7$ |
|---|---|---|---|
| lmNormal | 0.75 | 0.85 | 0.95 |
| lmLogormal | 0.35 | 0.45 | 0.55 |

Table B.15: Values of $\omega$ selected for *Nonlinear* model

| Models | $X_1 = 0.7$ | $X_1 = 0.9$ | $X_1 = 1.1$ |
|---|---|---|---|
| Nonlinear | 0.55 | 0.65 | 0.75 |

Table B.16: Values of $\omega$ selected for approximate confidence bands

| Models | lmNormal | lmLognormal | Nonlinear |
|---|---|---|---|
| $\omega$ | 1.00 | 0.60 | 0.80 |

## B.8 More details on the U.S. wage dataset

The data are extracted from U.S. 1980 1% metro sample from the Integrated Public Use Microdata Series (IPUMS) website (Ruggles et al., 2020). We first collect data of black and white people aged 30 - 60 with at least kindergarten level of education (nursery school is excluded), with positive annual earnings in the year preceding the census. Individuals with missing values for age, education and earnings are also excluded from the sample. Then we randomly sample $10,000$ people from the single and married groups, respectively and combine the resulting $20,000$ data as the final U.S. wage dataset.

The wage variable (wage) is the log annual wage, calculated as the log of the reported annual income from work in the previous year. The education variable (edu) corresponds to the highest grade of school completed starting from $0$ which indicates the kindergarten level. For the marital status variable (marital_status), 0 stands for "being single" and 1 stands for "being married". For the race variable (race), 1 corresponds to white people and 2 corresponds to black people. For the sex variable (sex), 1 corresponds to male and 2 correspond to female.

APPENDIX C

**APPENDIX OF CHAPTER 3**

## C.1 Technical tools

In this section, we collect technical tools that will be used in the subsequent proofs. For a probability measure $Q$ on a measurable space $(S, \mathcal{S})$ and a class of measurable functions $\mathcal{F}$ on $S$ such that $\mathcal{F} \subset L^2(Q)$, let $N(\mathcal{F}, \|\cdot\|_{Q,2}, \delta)$ denote the $\delta$-covering number for $\mathcal{F}$ with respect to the $L^2(Q)$-seminorm $\|\cdot\|_{Q,2}$. The class $\mathcal{F}$ is said to be pointwise measurable if there exists a countable subclass $\mathcal{G} \subset \mathcal{F}$ such that for every $f \in \mathcal{F}$ there exists a sequence $g_m \in \mathcal{G}$ with $g_m \to f$ pointwise. A function $F : S \to [0, \infty)$ is said to be an envelope for $\mathcal{F}$ if $F(x) \geq \sup_{f \in \mathcal{F}} |f(x)|$ for all $x \in S$. See Section 2.1 in van der Vaart and Wellner (1996) for details. For a vector-valued function $g$ defined over a set $T$, we define $\|g\|_T := \sup_{x \in T} \|g(x)\|$.

**Lemma 21** (Talagrand's concentration inequality)**.** *Let $\mathcal{F}$ be a pointwise measurable class of functions $S \mapsto \mathbb{R}$ and $X, X_1, \ldots, X_n$ be i.i.d. random variables following law $Q$. Suppose that there exists a constant $B > 0$ such that $\|f\|_\infty \leq B$ for all $f \in \mathcal{F}$. Suppose further that $\mathbb{E}f(X) = 0$ for all $f \in \mathcal{F}$. Let $\sigma^2 > 0$ be any positive constant such that $\sigma^2 \geq \sup_{f \in \mathcal{F}} \mathbb{E}f^2(X)$. Let $Z := \|\sum_{i=1}^n f(X_i)\|_{\mathcal{F}}$ and $V := n\sigma^2 + 2B\mathbb{E}[Z]$. Then for every $x > 0$,*

$$\mathbb{P}\big[Z \geq \inf_{\alpha > 0} \big\{(1+\alpha)\mathbb{E}[Z] + \sigma\sqrt{2nx} + (1/3 + 1/\alpha)Bx\big\}\big] \leq e^{-x}.$$

*In particular, we have*

$$\mathbb{P}\big[Z \geq 2\mathbb{E}[Z] + \sigma\sqrt{2nx} + 4/3Bx\big] \leq e^{-x}.$$

*Proof.* From Bousquet (2002), we have

$$\mathbb{P}\big[Z \geq \mathbb{E}[Z] + \sqrt{2Vx} + Bx/3\big] \leq \mathrm{e}^{-x}.$$

Then the first result follows from the observation that for any $\alpha > 0$,

$$\begin{aligned}\sqrt{2Vx} &\leq \sigma\sqrt{2nx} + 2\sqrt{Bx\mathbb{E}Z} \\ &\leq \sigma\sqrt{2nx} + \alpha\mathbb{E}Z + \alpha^{-1}Bx.\end{aligned}$$

The second result follows from taking $\alpha = 1$. □

## C.2 Proofs for Section 3.3.1

In the following, we define the augmented loss function on machine 1 at the $(t+1)$-th round

$$\tilde{\mathcal{L}}_1(\boldsymbol{\beta}; \hat{\boldsymbol{\beta}}_t) := \mathcal{L}_1(\boldsymbol{\beta}) + \left\langle \frac{1}{m}\sum_{i=1}^{m} \nabla\mathcal{L}_i(\hat{\boldsymbol{\beta}}_t) - \nabla\mathcal{L}_1(\hat{\boldsymbol{\beta}}_t), \boldsymbol{\beta} \right\rangle,$$

where $\hat{\boldsymbol{\beta}}_t$ is the estimator from the $t$-th round. For two sequences of positive constants $\{a_n\}_{n=1}^{\infty}$ and $\{b_n\}_{n=1}^{\infty}$, we denote $a_n \lesssim b_n$ if there exists a constant $C$ independent of both $n$ and iteration $t$ such that $a_n \leq Cb_n$. We use $C$, $C'$ and $C''$ to denote constants that are independent of both $n$ and iteration $t$ but may vary from line to line.

### C.2.1 Proof of Theorem 6

We will prove Theorem 6 in two steps. In the first step, we prove a recursive bound for the oracle estimator $\hat{\boldsymbol{\beta}}_{t+1}^{\circ}$; then in the second step, we show that $\hat{\boldsymbol{\beta}}_{t+1}$

coincides with $\hat{\boldsymbol{\beta}}^{\circ}_{t+1}$ with high probability. We finish the proof by a recursive argument based on the two steps. We assume without loss of generality that $\mathcal{S} = \{1, \ldots, s\}$.

**Step 1.** In this step, we will prove a recursive bound for $\|\hat{\boldsymbol{\beta}}^{\circ}_{t+1} - \boldsymbol{\beta}_0\|$ by relating it with that for $\|\hat{\boldsymbol{\beta}}_t - \boldsymbol{\beta}_0\|$. To this end, we provide the following recursive bound for the convergence rate of the oracle estimator.

**Lemma 22** (Recursive bound for oracle estimator)**.** *Suppose Assumption 4 holds and fix any $t \geq 0$. Pick any $\Delta_t \geq n^{-1} s \log n$, and assume that for all $1 \leq k \leq s$,*

$$\lambda_{t+1}\omega_{t+1,k} \leq \begin{cases} C\sqrt{\Delta_t s K_n^2 \log p/n} \vee \sqrt{K_n^2 \log n/N} & \text{if } t = 0 \\ C\sqrt{\Delta_t s K_n^2 \log n/n} \vee \sqrt{K_n^2 \log n/N} & \text{if } t \geq 1 \end{cases}$$

*Then, for any $\delta \in (0,1)$, we have that with probability at least $1 - \delta - Csn^{-s} - I\{t \geq 1\}\mathbb{P}(\|\hat{\boldsymbol{\beta}}_t - \boldsymbol{\beta}_0\| > \Delta_t \text{ or } \hat{\boldsymbol{\beta}}_t \neq \hat{\boldsymbol{\beta}}^{\circ}_t) - I\{t = 0\}\mathbb{P}(\|\hat{\boldsymbol{\beta}}_{ini} - \boldsymbol{\beta}_0\| > \Delta_0 \text{ or } \|\hat{\boldsymbol{\beta}}_{ini}\|_0 \geq Cs)$,*

$$\|\hat{\boldsymbol{\beta}}^{\circ}_{t+1} - \boldsymbol{\beta}_0\| \leq \begin{cases} C\Big(\sqrt{\frac{s^2 K_n^2 \log p}{n}} \cdot \Delta_t^{1/2} \bigvee \sqrt{\frac{s K_n^2 \log(n/\delta)}{N}}\Big) & \text{if } t = 0 \\ C\Big(\sqrt{\frac{s^2 K_n^2 \log n}{n}} \cdot \Delta_t^{1/2} \bigvee \sqrt{\frac{s K_n^2 \log(n/\delta)}{N}}\Big) & \text{if } t \geq 1 \end{cases}$$

*provided*

$$(I\{t = 0\}\sqrt{\log p} + I\{t \geq 1\}\sqrt{\log n})\sqrt{\frac{s^2 K_n^2}{n}} \cdot \Delta_t^{1/2} \bigvee \sqrt{\frac{s K_n^2 \log(n/\delta)}{N}} \leq C\frac{q\underline{f}\lambda_{\min}}{\bar{f}'\lambda_{\max}^{3/2}}.$$

The proof of Lemma 22 can be found in Appendix C.5.

Based on the above recursive bound and the initial conditions in Theorem 6, we can show, with probability at least $1 - Csn^{-s} - \mathbb{P}(\|\hat{\boldsymbol{\beta}}_{ini} - \boldsymbol{\beta}_0\| > C\eta_0 \text{ or } \|\hat{\boldsymbol{\beta}}_{ini}\|_0 \geq Cs)$

$$\|\hat{\boldsymbol{\beta}}^{\circ}_1 - \boldsymbol{\beta}_0\| \leq C\Big(\sqrt{\frac{s^2 K_n^2 \log p}{n}} \cdot \Delta_0^{1/2} \bigvee \sqrt{\frac{s K_n^2 \log n}{N}}\Big).$$

**Step 2.** In this step, we prove $\hat{\boldsymbol{\beta}}^{\circ}_{t+1}$ coincides with $\hat{\boldsymbol{\beta}}_{t+1}$ with high probability which is shown in the next Lemma.

**Lemma 23.** *Assume the conditions of Lemma 22 and further assume that* $(\min_{s+1\le k\le p}\omega_{t+1,k})^{-1}\Delta_t^{-1} = O(1)$ *and* $K_n\sqrt{s\log p}\Delta_t\big((a_n\sqrt{s}\ \vee\ 1)\sqrt{\Delta_t/n}\ \vee\ N^{-1/2}\big)\lambda_{t+1}^{-1} = o(1)$. *Then, for any* $\delta\in(0,1)$, *we have that with probability at least* $1-\delta-Cpn^{-s}-I\{t\ge 1\}\mathbb{P}(\|\hat{\boldsymbol{\beta}}_t-\boldsymbol{\beta}_0\|>\Delta_t$ *or* $\hat{\boldsymbol{\beta}}_t\neq\hat{\boldsymbol{\beta}}_t^\circ)-I\{t=0\}\mathbb{P}(\|\hat{\boldsymbol{\beta}}_0-\boldsymbol{\beta}_0\|>\Delta_0$ *or* $\|\hat{\boldsymbol{\beta}}_0\|_0\ge Cs)$,

$$\hat{\boldsymbol{\beta}}_{t+1}=\hat{\boldsymbol{\beta}}_{t+1}^\circ.$$

The proof of Lemma 23 can be found in Appendix C.5.

Combining Lemma 23 and the step 1, we have with probability at least $1-Cpn^{-s}-\mathbb{P}(\|\hat{\boldsymbol{\beta}}_{ini}-\boldsymbol{\beta}_0\|>C\eta_0$ or $\|\hat{\boldsymbol{\beta}}_{ini}\|_0\ge Cs)$,

$$\|\hat{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_0\|\le C\Big(\sqrt{\frac{s^2K_n^2\log p}{n}}\cdot\Delta_0^{1/2}\bigvee\sqrt{\frac{sK_n^2\log n}{N}}\Big).$$

Now we finish the proof by applying Lemmas 22 and 23 recursively. Specifically, we have for fixed $t\ge 0$, with probability at least $1-Ctpn^{-s}-\mathbb{P}(\|\hat{\boldsymbol{\beta}}_{ini}-\boldsymbol{\beta}_0\|>C\eta_0$ or $\|\hat{\boldsymbol{\beta}}_{ini}\|_0\ge Cs)$,

$$\Delta_t\le C\Big(\frac{s^2K_n^2\log n}{n}\cdot\Big(\sqrt{\frac{n}{K_n^2\log n}}\Big)^{1/2^t}\bigvee\sqrt{\frac{sK_n^2\log n}{N}}\Big),$$

which further implies that if we take $t\ge t_{opt}$ with probability at least $1-Ct_{opt}pn^{-s}-\mathbb{P}(\|\hat{\boldsymbol{\beta}}_{ini}-\boldsymbol{\beta}_0\|>C\eta_0$ or $\|\hat{\boldsymbol{\beta}}_{ini}\|_0\ge Cs)$,

$$\eta_t=O\big(\sqrt{sK_n^2\log n/N}\big).$$

We finish the proof by noting $\mathbb{P}(\|\hat{\boldsymbol{\beta}}_{ini}-\boldsymbol{\beta}_0\|>C\eta_0$ or $\|\hat{\boldsymbol{\beta}}_{ini}\|_0\ge Cs)\le\gamma_n$ under our assumptions. □

### C.2.2 Proof of Corollary 2

The main proof follows the same recursive argument as the proof of Theorem 6 except that we need to show at each iteration, the conditions involving $\omega$ in Assumption 5 can hold under the new choice of $\omega$ with high probability.

First, for the first iteration, i.e., $t = 1$, the conditions of $\hat{\beta}_{\text{ini}}$ together with condition (i) imply with probability at least $1 - \gamma_n$, $\max_{k\in\mathcal{S}} \lambda_{t+1}|\hat{\beta}_{\text{ini},k}|^{-1} = O\big(\sqrt{\eta_0 s K_n^2 \log n/n} \vee \sqrt{K_n^2 \log n/N}\big)$. Combining with condition (ii), we can show the conditions involving $\omega$ in Assumption 5 hold with probability at least $1 - \gamma_n$ which finishes the proof of the current case.

For iteration $t + 1$ ($t >= 1$), by the proof of Theorem 6, we have shown with probability at least $1 - C(t-1)pn^{-s} - \gamma_n$, we have

$$\hat{\boldsymbol{\beta}}_t = \hat{\boldsymbol{\beta}}_t^{\circ} \quad \text{and} \quad \|\hat{\boldsymbol{\beta}}_t^{\circ} - \boldsymbol{\beta}_0\| \leq C\eta_t,$$

which, under condition (i), further implies

$$\max_{k\in\mathcal{S}} \lambda_{t+1}|\hat{\beta}_{t,k}|^{-1} = O\big(\sqrt{\eta_t s K_n^2 \log n/n} \vee \sqrt{K_n^2 \log n/N}\big).$$

Given the above result, it is easy to see the conditions involving $\omega$ in Assumption 5 hold with probability at least $1 - C(t-1)pn^{-s} - \gamma_n$ under the new choices of $\omega_{t+1,k}$ and $\lambda_{t+1}$. The rest of the proof follows the same argument as Theorem 6 and is omitted for brevity.

### C.2.3 Proof of Theorem 7

The proof is divided into two parts. In the first step, we prove the convergence rate in the conclusion based on a recursive argument. In the second step, we

show the sparsity conclusion of the theorem. We introduce some notations. Define event $E_1 := \{0 < c_1' \leq \bar{\phi}_{\min}(\ell_n s, \boldsymbol{\Sigma}_n) \leq \bar{\phi}_{\max}(\ell_n s, \boldsymbol{\Sigma}_n) \leq C_1'\}$ and $E_2 := \big\{\|n^{-1}\sum_{j=1}^n \boldsymbol{X}_{1j}\boldsymbol{X}_{1j}^T - \boldsymbol{\Sigma}\|_\infty > CK_n^2\sqrt{\frac{\log(p/\delta)}{n}}$ or $\|(mn)^{-1}\sum_{i=1}^m\sum_{j=1}^n \boldsymbol{X}_{ij}\boldsymbol{X}_{ij}^T - \boldsymbol{\Sigma}\|_\infty > CK_n^2\sqrt{\frac{\log(p/\delta)}{N}}$ or $\big\|\frac{1}{m}\sum_{i=1}^m \nabla\mathcal{L}_i(\boldsymbol{\beta}_0)\big\|_\infty > CV_nK_n\sqrt{\frac{\log(p/\delta)}{N}}\big\}$ for some constant $\delta > 0$ independent of $n$ and $t$. Recall, we have $\mathbb{P}(E_1) \geq 1 - \nu_n$.

**Step 1.** We will show the convergence rate of the estimator by a recursive argument similar to the first step in the proof of Theorem 6. The following lemma provides the recursive bound which will be fundamental to establish the convergence rate.

**Lemma 24** (Recursive bound for heteroscedastic lasso)**.** *Suppose Assumption 6 holds and fix any $t \geq 0$. We choose $\lambda_{t+1}$, satisfies $C_1\Gamma_t \geq \lambda_{t+1} \geq C_2\Gamma_t$ for some sufficiently large constants $C_1 \geq C_2$ independent of $n$ and $t$, and we define*

$$\Gamma_t := V_nK_n\sqrt{\frac{\log(p/\delta)}{N}} + K_n^2\sqrt{\frac{\log(p/\delta)}{n}}\cdot\Delta_t.$$

*Then with probability $1 - \mathbb{P}(E_2) - P(E_1^c) - \mathbb{P}(\|\hat{\boldsymbol{\beta}}_t - \boldsymbol{\beta}_0\|_1 > \Delta_t)$, we have*

$$\|\hat{\boldsymbol{\beta}}_{t+1} - \boldsymbol{\beta}_0\|_1 \leq C\Big[V_nK_ns\sqrt{\frac{\log(p/\delta)}{N}} + sK_n^2\sqrt{\frac{\log(p/\delta)}{n}}\cdot\Delta_t\Big],$$
$$\|\hat{\boldsymbol{\beta}}_{t+1} - \boldsymbol{\beta}_0\|_2 \leq C\Big[V_nK_n\sqrt{\frac{s\log(p/\delta)}{N}} + K_n^2\sqrt{\frac{s\log(p/\delta)}{n}}\cdot\Delta_t\Big].$$

*In particular, we have $\mathbb{P}(E_2) \leq \delta$.*

The proof of Lemma 24 can be found in Appendix C.5.

Now we prove the convergence rate in the theorem. By applying Lemma 24 recursively, with probability $1 - p^{-1} - \gamma_n - \nu_n$ (we choose $\delta = p^{-1}$),

$$\|\hat{\boldsymbol{\beta}}_t - \boldsymbol{\beta}_0\|_1 \leq (Ca)^tO(\eta_0) + \frac{1-(Ca)^t}{1-(Ca)}(Cb).$$

Hence, when $t \geq t_{\text{opt}}$, we have $(Ca)^t O(\eta_0) \leq Cb$ which implies

$$\|\hat{\boldsymbol{\beta}}_t - \boldsymbol{\beta}_0\|_1 \leq C V_n K_n s \sqrt{\frac{\log p}{N}}. \tag{C.1}$$

The convergence rate in $\ell_2-$norm can be shown by combining the recursive $\ell_2$ bound in Lemma 24 and $(C.1)$. This finishes the proof of the first step.

**Step 2.** We show the sparsity property of the estimator. To this end, let $\hat{\mathcal{S}}_{t+1}$ denote the support of $\hat{\boldsymbol{\beta}}_{t+1}$ and we define $\hat{m}_{t+1} := |\hat{\mathcal{S}}_{t+1} \backslash \mathcal{S}|$, i.e., the cardinality of $\hat{\mathcal{S}}_{t+1} \backslash \mathcal{S}$. We follow the proof approach of Lemma 10 in Belloni et al. (2012). The following empirical pre-sparsity lemma will be useful.

**Lemma 25** (Empirical pre-sparsity)**.** *Under the condition of Theorem 7, with probability* $1 - p^{-1} - \gamma_n - \nu_n$*,*

$$\hat{m}_{t+1} \leq C \phi_{\max}(\hat{m}_{t+1}, \boldsymbol{\Sigma}_n) s.$$

The proof of Lemma 25 can be found in Appendix C.5.

Now, we define set

$$\mathcal{M} := \{m \in \mathbb{N} : m > 2C\phi_{\max}(m)s\}.$$

Then, based on the pre-sparsity bound, an almost the same analysis as in Lemma 10 of Belloni et al. (2012) yields, with the same probability as in Lemma 25, we have

$$\hat{m}_{t+1} \leq 2Cs\big(\min_{m \in \mathcal{M}} \phi_{\max}(m \wedge n, \boldsymbol{\Sigma}_n)\big). \tag{C.2}$$

Combining $(C.2)$ and conditions (ii) and (iii) in Assumption 6, we finally obtain $\hat{m}_{t+1} = O(s)$ with probability $1 - p^{-1} - \gamma_n - \nu_n$ which further implies the desired sparsity property of the conclusion. □

## C.3 Proofs for Section 3.3.2

### Proof of Proposition 3

To prove the proposition, we will need the following two lemmas which concern a stochastic expansion and the convergence rate of the kernel smoothing density estimator, respectively.

**Lemma 26** (Stochastic expansion)**.** *Under conditions of Proposition 3, the following stochastic expansion holds with probability* $1 - o(1)$*,*

$$\mathbb{E}_N[\varphi(y_j - \boldsymbol{X}_j^T\hat{\boldsymbol{\beta}}_0)(\boldsymbol{X}_{j1} - \boldsymbol{X}_{j(-1)}^T\hat{\boldsymbol{\theta}}_0)] = -\big(f_\epsilon\mathbb{E}[v^2]\big)(\hat{\alpha}_1 - \alpha_1) + \mathbb{E}_N[\varphi(\epsilon_j)v_j] + R_2,$$

*where residue term* $R_2 = O\big(c_n(s\log a_n/N)^{3/4}\big)$*.*

**Lemma 27** (Density estimation)**.** *Under conditions of Proposition 3, we have*

$$|\hat{f}_\epsilon(0) - f_\epsilon(0)| = O_P\big(h^2 + (Nh)^{-1/2} + N^{-1/2}h^{-2}\rho_N\big).$$

The proof of Lemmas 26 and 27 can be found in Appendix C.5.

Now, we prove Proposition 3. First, we note

$$\begin{aligned}|\mathbb{E}_N[\hat{v}_j^2] - \mathbb{E}[v^2]| &\le |\mathbb{E}_N[\hat{v}_j^2] - \mathbb{E}_N[v_j^2]| + |\mathbb{E}_N[v_j^2] - \mathbb{E}[v^2]| \\ &= O_P(\rho_N) + O_P(N^{-1/2}) \\ &= O_P(\rho_N).\end{aligned}$$

Also by Lemma 27, we have

$$|\hat{f}_\epsilon(0) - f_\epsilon(0)| = O_P\big(h^2 + (Nh)^{-1/2} + N^{-1/2}h^{-2}\rho_N\big).$$

Combining above two results and choosing $h = O(\rho_N^{1/2})$, we can show

$$\hat{f}_\epsilon(0)\mathbb{E}_N[\hat{v}_j^2] = f_\epsilon(0)\mathbb{E}[v^2] + R_1, \tag{C.3}$$

where

$$R_1 = O_P(\rho_N \vee c_n^{-1/4} N^{-3/8} (s \log a_n)^{-1/8}).$$

By Lemma 26,

$$\mathbb{E}_N[\varphi(y_j - \boldsymbol{X}_j^T \hat{\boldsymbol{\beta}}_0)(\boldsymbol{X}_{j1} - \boldsymbol{X}_{j(-1)}^T \hat{\boldsymbol{\theta}}_0)] = -\big(f_\epsilon(0)\mathbb{E}[v^2]\big)(\hat{\alpha}_1 - \alpha_1) + \mathbb{E}_N[\varphi(\epsilon_j)v_j] + R_2, \tag{C.4}$$

where

$$R_2 = O_P(c_n (s \log a_n / N)^{3/4}).$$

Now combining (C.3) and (C.4), we can show

$$\begin{aligned}
\check{\alpha}_1 &= \alpha_1 + \big(f_\epsilon(0)\mathbb{E}[v^2]\big)^{-1} \cdot \mathbb{E}_N[\varphi(U_j - 1/2)v_j] + O\big(R_2 - R_1\rho_n\big) \\
&= \alpha_1 + \big(f_\epsilon(0)\mathbb{E}[v^2]\big)^{-1} \cdot \mathbb{E}_N[\varphi(U_j - 1/2)v_j] + O_P\big(c_n (s \log a_n / N)^{3/4}\big),
\end{aligned}$$

where we used the fact $\varphi(U_j - 1/2)$ has the same distribution as $\varphi(\epsilon_j)$ and this finishes the proof. □

## C.4 Proofs for Section 2.3.3

### Proof of Proposition 4

The proof is almost identical to Proposition 3 but here we require the convergence rates in the proof to be uniform in coordinates of interest.

First, we note,

$$\begin{aligned}\sup_{1\le k\le p_1}|\mathbb{E}_N[\hat{v}^2_{(k)j}]-\mathbb{E}[v^2_{(k)}]| &\le \sup_{1\le k\le p_1}|\mathbb{E}_N[\hat{v}^2_{(k)j}]-\mathbb{E}_N[v^2_{(k)j}]|+\sup_{1\le k\le p_1}|\mathbb{E}_N[v^2_{(k)j}]-\mathbb{E}[v^2_{(k)}]|\\ &= O_P(\rho_N)+O_P(V_n\sqrt{\log p_1/N})\\ &= O_P(\rho_N).\end{aligned}$$

Also by Lemma 27 and the choice of bandwidth $h$, we have

$$|\hat{f}_\epsilon(0)-f_\epsilon(0)| = O_P\big(\rho_N \vee c_n^{-1/4}N^{-3/8}(s\log a_n)^{-1/8}\big).$$

Combining above two results we can show, uniform in $1\le k\le p_1$,

$$\hat{f}_\epsilon(0)\mathbb{E}_N[\hat{v}^2_{(k)j}] = f_\epsilon(0)\mathbb{E}[v^2_{(k)}]+R_1, \tag{C.5}$$

where $R_1 = O_P\big(\rho_N \vee c_n^{-1/4}N^{-3/8}(s\log a_n)^{-1/8}\big)$.

It is a routine to verify the stochastic expansion in Lemma 26 holds uniformly in coordinates under assumptions of Proposition 4, i.e., uniform in $1\le k\le p_1$,

$$\mathbb{E}_N[\varphi(y_j-\boldsymbol{X}_j^T\hat{\boldsymbol{\beta}}_0)\hat{v}_{(k)j}] = -\big(f_\epsilon(0)\mathbb{E}[v^2_{(k)}]\big)(\hat{\alpha}_k-\alpha_k)+\mathbb{E}_N[\varphi(U_j-1/2)v_{(k)j}]+R_2, \tag{C.6}$$

where by the growth condition of the theorem,

$$R_2 = O_P(c_n(s\log a_n/N)^{3/4}).$$

Now combining (C.5) and (C.6), we can show, uniform in $1\le k\le p_1$

$$\begin{aligned}\check{\alpha}_k &= \alpha_k+[f_\epsilon(0)\mathbb{E}(v^2_{(k)})]^{-1}\cdot\mathbb{E}_N[\varphi(U_j-1/2)v_{(k)j}]+O\big(R_2-R_1\rho_N\big)\\ &= \alpha_k+[f_\epsilon(0)\mathbb{E}(v^2_{(k)})]^{-1}\cdot\mathbb{E}_N[\varphi(U_j-1/2)v_{(k)j}]+O_P\big(c_n(s\log a_n/N)^{3/4}\big),\end{aligned}$$

which finishes the proof. □

## Proof of Theorem 8

We divide the proof into two steps.

**Step 1.** We show

$$\sup_{\boldsymbol{b}\in\mathbb{R}^{p_1}} \left|\mathbb{P}\big(N^{-1/2}\textstyle\sum_{j=1}^{N} \boldsymbol{A}\Psi_j \le \boldsymbol{b}\big) - \mathbb{P}\big(\boldsymbol{AG} \le \boldsymbol{b}\big)\right| \to 0.$$

This will be down by verifying conditions (M.1), (M.2) and (E.2) of Proposition 2.1 in Chernozhukov et al. (2017a).

**(M.1)**: For $1 \le k \le p_1$,

$$N^{-1}\sum_{j=1}^{N}(4f_\epsilon^2(0)\mathbb{E}[v_{(k)}^2]) \cdot \mathbb{E}\big[\psi^2(U_j, v_{(k)j})\big] = 1.$$

**(M.2)**: First we note $\big(\boldsymbol{A}\Psi_j\big)_k = 2\mathbb{E}[v_{(k)}^2]^{-1/2}\varphi(U_j - 1/2)v_{(k)j}$. Hence, for $1 \le k \le p_1$,

$$\begin{aligned}
&N^{-1}\sum_{j=1}^{N}\mathbb{E}\big[2^3\mathbb{E}[v_{(k)}^2]^{-3/2} \cdot |1/2 - I\{U_j \le 0\}|^3 \cdot |v_{(k)j}|^3\big] \\
&= \mathbb{E}[v_{(k)}^2]^{-3/2} \cdot \mathbb{E}[|v_{(k)}|^3] \le Const.
\end{aligned}$$

Similarly, we can show

$$N^{-1}\sum_{j=1}^{N}\mathbb{E}\big[2^4\mathbb{E}[v_{(k)}^2]^{-2} \cdot |1/2 - I\{U_j \le 0\}|^4 \cdot |v_{(k)j}|^4\big] \le Const.$$

**(E.2)** For $1 \le j \le n$ and any $q \ge 4$

$$\begin{aligned}
&\mathbb{E}\Big[\big(\max_{1\le k\le p_1} 2\mathbb{E}[v_{(k)}^2]^{-1/2}|\varphi(U_j - 1/2)v_{(k)j}|\big)^q\Big] \\
\le&(2/\sqrt{c_1})^q \cdot (1/2)^q \cdot \mathbb{E}\Big[\big(\max_{1\le k\le p_1}|v_{(k)j}|\big)^q\Big] \\
\le&c_1^{-q/2}V_n^q.
\end{aligned}$$

Hence by Proposition 2.1 in Chernozhukov et al. (2017a), the conclusion of this step follows as soon as

$$\frac{V_n^2 \log^7(p_1 n)}{n} \to 0,$$

which holds under our assumption.

**Step 2.** Define $\boldsymbol{R} := (\check{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) - N^{-1} \sum_{j=1}^N \Psi_j$ and by Proposition 4, $||\boldsymbol{R}||_\infty = O_P\big(c_n(s \log a_n/N)^{3/4}\big)$. This further implies

$$\|\sqrt{n}\boldsymbol{A}\boldsymbol{R}\|_\infty = O_P\big(c_n(s \log a_n)^{3/4} N^{-1/4}\big).$$

Define $\delta_n := c_n(s \log a_n)^{3/4} N^{-1/4}$ and we have

$$\mathbb{P}\big(\|\sqrt{n}\boldsymbol{A}\boldsymbol{R}\|_\infty \geq B_n \delta_n\big) \to 0,$$

for any sequence of constants $B_n \to \infty$. Now, for any $\boldsymbol{b} \in \mathbb{R}^{p_1}$,

$$\begin{aligned}
\mathbb{P}\big(\sqrt{N}\boldsymbol{A}(\check{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) \leq \boldsymbol{b}\big) &\leq \mathbb{P}\big(N^{-1/2} \textstyle\sum_{j=1}^N \boldsymbol{A}\Psi_j \leq \boldsymbol{b} + B_n\delta_n\big) + o(1) \\
&\leq \mathbb{P}\big(\boldsymbol{A}\boldsymbol{G} \leq \boldsymbol{b} + B_n\delta_n\big) + o(1) \quad \text{(by step 1)} \\
&\leq \mathbb{P}\big(\boldsymbol{A}\boldsymbol{G} \leq \boldsymbol{b}\big) + O\big(B_n\delta_n\sqrt{\log p_1}\big) + o(1) \quad \text{(by Nazarovs inequality)}.
\end{aligned}$$

Similarly, we can show

$$\mathbb{P}\big(\sqrt{n}\boldsymbol{A}(\check{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) \leq \boldsymbol{b}\big) \geq \mathbb{P}\big(\boldsymbol{A}\boldsymbol{G} \leq \boldsymbol{b}\big) - O\big(B_n\delta_n\sqrt{\log p_1}\big) - o(1).$$

Since $B_n\delta_n\sqrt{\log p_1} \to 0$ under our assumption for $B_n$ grows sufficiently slow, we proved the theorem.

## Proof of Theorem 9

Define $\hat{\boldsymbol{\Sigma}}_1 := N^{-1}\sum_{j=1}^N \mathbb{E}_U\big[\hat{\Psi}_j\hat{\Psi}_j^T\big]$. We note

$$
\begin{aligned}
&\sup_{\boldsymbol{b}\in\mathbb{R}^{p_1}} \big|\mathbb{P}_U\big(N^{-1/2}\textstyle\sum_{j=1}^N \hat{\boldsymbol{A}}\hat{\Psi}_j \le \boldsymbol{b}\big) - \mathbb{P}\big(\boldsymbol{AG}\le \boldsymbol{b}\big)\big| \\
\le &\sup_{\boldsymbol{b}\in\mathbb{R}^{p_1}} \big|\mathbb{P}_U\big(N^{-1/2}\textstyle\sum_{j=1}^n \hat{\boldsymbol{A}}\hat{\Psi}_j \le \boldsymbol{b}\big) - \mathbb{P}_U\big(\hat{\boldsymbol{A}}\hat{\boldsymbol{G}}\le \boldsymbol{b}\big)\big| \\
&+ \sup_{\boldsymbol{b}\in\mathbb{R}^{p_1}} \big|\mathbb{P}_U\big(\hat{\boldsymbol{A}}\hat{\boldsymbol{G}}\le \boldsymbol{b}\big) - \mathbb{P}_U\big(\boldsymbol{A}\hat{\boldsymbol{G}}\le \boldsymbol{b}\big)\big| + \sup_{\boldsymbol{b}\in\mathbb{R}^{p_1}} \big|\mathbb{P}_U\big(\boldsymbol{A}\hat{\boldsymbol{G}}\le \boldsymbol{b}\big) - \mathbb{P}\big(\boldsymbol{AG}\le \boldsymbol{b}\big)\big| \\
=:&I + II + III,
\end{aligned}
$$

where $\hat{\boldsymbol{G}} \sim N(0,\hat{\boldsymbol{\Sigma}}_1)$ with $\hat{\boldsymbol{\Sigma}}_1 := N^{-1}\sum_{j=1}^N \mathbb{E}_U\big[\hat{\Psi}_j\hat{\Psi}_j^T\big]$. Now we analyze the three terms respectively.

To show $I = o_P(1)$, we apply Proposition 2.1 in Chernozhukov et al. (2017a) conditioning on the data and verify conditions (M.1), (M.2) and (E.2). We can verify (M.1) by noting

$$
N^{-1}\sum_{j=1}^N (4\hat{f}_\epsilon^2(0)\mathbb{E}_N[\hat{v}_{(k)j}^2])\cdot \mathbb{E}_U\big[\hat{\psi}^2(U_j,\hat{v}_{(k)j})\big] = 1.
$$

By a similar analysis in Theorem 8, we can show, for $1\le k\le p_1$ and $l = 3$ or $4$,

$$
N^{-1}\sum_{j=1}^N \mathbb{E}_U\big[2^l\mathbb{E}_N[\hat{v}_{(k)j}^2]^{-l/2}\cdot|1/2 - I\{U_j\le 0\}|^l\cdot|\hat{v}_{(k)j}|^l\big] \le Const,
$$

with probability approaching 1, where the constant on the right hand side is independent of $n$. Finally, for any $q\ge 4$ and $1\le j\le N$,

$$
\begin{aligned}
&\mathbb{E}_U\Big[\big(\max_{1\le k\le p_1} 2\mathbb{E}_N[\hat{v}_{(k)j}^2]^{-1/2}|\varphi(U_j-1/2)\hat{v}_{(k)j}|\big)^q\Big] \\
\le\ & c_1^{-q/2}\Big(\max_{1\le k\le p_1}|\hat{v}_{(k)j}|^q\Big) \\
\le\ & c_1^{-q/2}\cdot O\big((V_n+\rho_N)^q\big).
\end{aligned}
$$

Hence by Proposition 2.1 in Chernozhukov et al. (2017a), we can see $I = o_P(1)$ as soon as

$$\frac{(V_n + \rho_N)^2 \log^7(p_1 N)}{N} \to 0,$$

which holds under our assumption.

To show $II = o_P(1)$, we invoke the Gaussian comparison lemma (Lemma 4). Hence it suffices to show

$$\left\| \hat{\boldsymbol{A}} \hat{\boldsymbol{\Sigma}}_1 \hat{\boldsymbol{A}} - \boldsymbol{A} \hat{\boldsymbol{\Sigma}}_1 A \right\|_\infty \cdot \log^2 p_1 \to 0.$$

In fact, by a similar analysis as in Lemma 27,

$$\begin{aligned}
\left\| \hat{\boldsymbol{A}} \hat{\boldsymbol{\Sigma}}_1 \hat{\boldsymbol{A}} - \boldsymbol{A} \hat{\boldsymbol{\Sigma}}_1 \boldsymbol{A} \right\|_\infty &= \max_{1 \le j,k \le p_1} |(\hat{a}_j \hat{a}_k - a_j a_k) \hat{\boldsymbol{\Sigma}}_{1,jk}| \\
&= O_P\big(\rho_N \vee c_n^{-1/4} N^{-3/8} (s \log a_n)^{-1/8}\big) \cdot O_P(1) \\
&= O_P\big(\rho_N \vee c_n^{-1/4} N^{-3/8} (s \log a_n)^{-1/8}\big).
\end{aligned}$$

Hence, in view of the Gaussian comparison lemma, we have $II = o_P(1)$ as soon as

$$\big(\rho_N \vee c_n^{-1/4} N^{-3/8} (s \log a_n)^{-1/8}\big) \log^2 p_1 \to 0,$$

which holds under our assumption.

We again apply Gaussian comparison to show $III = o_P(1)$. First, by a similar analysis as in Lemma 27, we can show

$$\left\| \hat{\boldsymbol{\Sigma}}_1 - \boldsymbol{\Sigma}_1 \right\|_\infty = O_P\big(\rho_N \vee c_n^{-1/4} N^{-3/8} (s \log a_n)^{-1/8}\big).$$

This further implies

$$\left\| \boldsymbol{A} \hat{\boldsymbol{\Sigma}}_1 \boldsymbol{A} - \boldsymbol{A} \boldsymbol{\Sigma}_1 \boldsymbol{A} \right\|_\infty = O_P\big(\rho_N \vee c_n^{-1/4} N^{-3/8} (s \log a_n)^{-1/8}\big).$$

Now, combining the Gaussian comparison lemma and the growth condition in the theorem, we have $III = o_P(1)$. This completes the proof.

## Proof of Theorem 10

We follow the same two steps as Theorem 8. The Step 1 is exactly the same and we only point out the difference in the Step 2 here.

We note Theorem 1 implies

$$\sup_{\boldsymbol{b}\in\mathbb{R}^{p_1}} \left|\mathbb{P}\big(\sqrt{N}(\check{\boldsymbol{\alpha}}-\boldsymbol{\alpha})\le \boldsymbol{b}\big) - \mathbb{P}\big(\boldsymbol{G}\le \boldsymbol{b}\big)\right| \to 0.$$

Additionally, since the variances of the coordinates of $\boldsymbol{G}$ are bounded, we see that $\mathbb{E}[\|\boldsymbol{G}\|_\infty] = O(\sqrt{\log p_1})$ by Lemma 2.2.2 in van der Vaart and Wellner (1996). Hence, we have

$$\|\sqrt{n}(\check{\boldsymbol{\alpha}}-\boldsymbol{\alpha})\|_\infty = O_P(\sqrt{\log p_1}).$$

Also as in the proof of Theorem 9, we have

$$\|\hat{\boldsymbol{A}}-\boldsymbol{A}\|_\infty = O_P\big(\rho_N \vee c_n^{-1/4}N^{-3/8}(s\log a_n)^{-1/8}\big).$$

Combining the above rates with the growth condition of the theorem, we have shown

$$\big\|(\hat{\boldsymbol{A}}-\boldsymbol{A})\sqrt{N}(\check{\boldsymbol{\alpha}}-\boldsymbol{\alpha})\big\|_\infty = o_P(1/\sqrt{\log p_1}).$$

The rest of the proof is analogous to the last part of Theorem 8.

## C.5 Proofs for Auxiliary Results

## Proof of Lemma 22

The following lemma which provides a bound on the $\ell^\infty$-norm of the subgradient of the loss function will be important for the proof.

**Lemma 28.** *Suppose Assumption 4 holds and fix any $t \geq 0$. Pick any $\Delta_t \geq n^{-1}s\log n$ and $\delta \in (0,1)$. Then with probability at least $1-\delta-Csn^{-s}-I\{t\geq 1\}\mathbb{P}(\|\hat{\boldsymbol{\beta}}_t-\boldsymbol{\beta}_0\|>\Delta_t$ or $\hat{\boldsymbol{\beta}}_t \neq \hat{\boldsymbol{\beta}}_t^\circ)-I\{t=0\}\mathbb{P}(\|\hat{\boldsymbol{\beta}}_{ini}-\boldsymbol{\beta}_0\|>\Delta_0$ or $\|\hat{\boldsymbol{\beta}}_{ini}\|_0 \geq Cs)$,*

$$\|\nabla\tilde{\mathcal{L}}_1(\boldsymbol{\beta}_0;\hat{\boldsymbol{\beta}}_t)_{\mathcal{S}}\|_\infty \leq \begin{cases} C\left[\sqrt{\frac{K_n^2\log(n/\delta)}{N}}+\sqrt{\frac{\Delta_t s K_n^2\log p}{n}}\right] & \text{if } t=0 \\ \\ C\left[\sqrt{\frac{K_n^2\log(n/\delta)}{N}}+\sqrt{\frac{\Delta_t s K_n^2\log n}{n}}\right] & \text{if } t\geq 1 \end{cases} \tag{C.7}$$

*Proof of Lemma 28.* In the following proof, we will omit the subscript $\mathcal{S}$ for brevity.

We first prove the case $t \geq 1$ and we condition on the event $\{\|\hat{\boldsymbol{\beta}}_t-\boldsymbol{\beta}_0\| \leq \Delta_t$ and $\hat{\boldsymbol{\beta}}_t=\hat{\boldsymbol{\beta}}_t^\circ\}$ which has probability at least $1-\mathbb{P}(\|\hat{\boldsymbol{\beta}}_t-\boldsymbol{\beta}_0\|>\Delta_t$ or $\hat{\boldsymbol{\beta}}_t \neq \hat{\boldsymbol{\beta}}_t^\circ)$. First, we have

$$\begin{aligned}
&\|\nabla\tilde{\mathcal{L}}_1(\boldsymbol{\beta}_0,\hat{\boldsymbol{\beta}}_t)\|_\infty = \|\nabla\mathcal{L}_1(\boldsymbol{\beta}_0)-\nabla\mathcal{L}_1(\hat{\boldsymbol{\beta}}_t)+m^{-1}\textstyle\sum_{i=1}^m\nabla\mathcal{L}_i(\hat{\boldsymbol{\beta}}_t)\|_\infty \\
&\leq \left\|m^{-1}\textstyle\sum_{i=1}^m\nabla\mathcal{L}_i(\boldsymbol{\beta}_0)\right\|_\infty + \left\|\nabla\mathcal{L}_1(\hat{\boldsymbol{\beta}}_t)-\nabla\mathcal{L}_1(\boldsymbol{\beta}_0)-\mathbb{E}[\{\tau-I(y\leq\boldsymbol{X}^T\boldsymbol{\beta})\}\boldsymbol{X}]|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_t}\right\|_\infty \\
&\quad + \left\|m^{-1}\textstyle\sum_{i=1}^m(\nabla\mathcal{L}_i(\hat{\boldsymbol{\beta}}_t)-\nabla\mathcal{L}_i(\boldsymbol{\beta}_0))-\mathbb{E}[\{\tau-I(y\leq\boldsymbol{X}^T\boldsymbol{\beta})\}\boldsymbol{X}]|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_t}\right\|_\infty \\
&:= I+II+III
\end{aligned}$$

We will analyze terms $I, II$, and $III$ separately.

**Terms II and III**. By union bound, we have

$$\begin{aligned}
&\mathbb{P}\left(II>u\right) \\
&\leq s\max_{1\leq k\leq s}\mathbb{P}\Bigg(\Bigg|\frac{1}{n}\sum_{j=1}^n\{I(y_{1j}\leq\boldsymbol{X}_{1j}^T\hat{\boldsymbol{\beta}}_t)-I(y_{1j}\leq\boldsymbol{X}_{1j}^T\boldsymbol{\beta}_0)\}X_{1jk}- \\
&\qquad\qquad \mathbb{E}[\{\tau-I(y\leq\boldsymbol{X}^T\boldsymbol{\beta})\}X_k]|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_t}\Bigg|>u\Bigg) \\
&:= s\max_{1\leq k\leq s}p_k.
\end{aligned}$$

We introduce the following function class, for $1 \le k \le s$,

$$\mathcal{F}_k := \{(\boldsymbol{x}, y) \mapsto \{I(y \le \boldsymbol{x}^T\boldsymbol{\beta}) - I(y \le \boldsymbol{x}^T\boldsymbol{\beta}_0)\}x_k : \ \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| \le \Delta_t, \mathcal{S}_{\boldsymbol{\beta}} \subset \mathcal{S}\}.$$

Then we have

$$p_k \le \mathbb{P}\left(n^{-1/2}\|\mathbb{G}_n\|_{\mathcal{F}_k} > u\right).$$

To bound the probability on the right hand side, we will first use the local maximal inequality (Lemma 2)to bound $\mathbb{E}[\|\mathbb{G}_n\|_{\mathcal{F}_k}]$ and then apply Talagrand's inequality (Lemma 21).

First, in view of the sparsity in the definition of $\mathcal{F}_k$, it is not difficult to see that $\mathcal{F}$ is VC-type with VC index $V \le Cs$. Applying Lemma 2 with $\sigma^2 = C\Delta_t K_n^2$ and $F = K_n$, we have

$$\mathbb{E}[\|\mathbb{G}_n\|_{\mathcal{F}_k}] \le C(\sqrt{\Delta_t K_n^2 s \log n} + n^{-1/2} s K_n \log n) \le C\sqrt{\Delta_t K_n^2 s \log n}, \qquad \text{(C.8)}$$

where we used $\Delta_t \ge n^{-1} s \log n$ in the last inequality.

Now we apply Talagrand's inequality (Lemma 21) to get

$$\mathbb{P}\left(\|\mathbb{G}_n\|_{\mathcal{F}_k} > 2\mathbb{E}\big[\|\mathbb{G}_n\|_{\mathcal{F}_k}\big] + \sqrt{2\bar{f}\lambda_{\max}\eta_t K_n^2} \cdot \sqrt{x} + 8/3 K_n n^{-1/2} x\right) \le \mathrm{e}^{-x}. \quad \text{(C.9)}$$

By plugging (C.8) into (C.9) and choosing $x = s \log n$, we have shown that with probability $1 - n^{-s}$,

$$n^{-1/2}\|\mathbb{G}\|_{\mathcal{F}_k} \le C\sqrt{n^{-1}\Delta_t K_n^2 s \log n}.$$

Conclude that with probability at least $1 - sn^{-s}$,

$$II \le C\sqrt{n^{-1}\Delta_t K_n^2 s \log n}.$$

Likewise, we have that with probability at least $1 - sn^{-ms}$,

$$III \le C\sqrt{n^{-1}\Delta_t K_n^2 s \log n}.$$

**Term I**. By Hoeffding's inequality together with union bound, we have

$$\begin{aligned}\mathbb{P}\left(I > u\right) &\leq s \max_{1\leq k\leq s} \mathbb{P}\left(\left|\frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\{\tau - I(y_{ij} \leq \boldsymbol{X}_{ij}^T\boldsymbol{\beta}_0)\}X_{ijk}\right| > u\right) \\ &\leq 2s\exp\Big(-\frac{Nu^2}{4K_n^2}\Big). \quad (N = mn)\end{aligned}$$

Thus $I \leq C\sqrt{K_n^2 \log(n/\delta)/N}$ with probability at least $1-\delta$.

The conclusion of this lemma for $t \geq 1$ follows from combining these bounds.

For $t = 0$, we follow the same analysis as above and proceed to analyze terms $I$, $II$ and $III$ conditioning on the event $\{\|\hat{\boldsymbol{\beta}}_{ini} - \boldsymbol{\beta}_0\| \leq \Delta_0$ and $\|\hat{\boldsymbol{\beta}}_{ini}\|_0 \leq Cs\}$ which has probability at least $1 - \mathbb{P}(\|\hat{\boldsymbol{\beta}}_{ini} - \boldsymbol{\beta}_0\| > \Delta_0$ or $\|\hat{\boldsymbol{\beta}}_{ini}\|_0 \geq Cs)$.

To bound terms $II$ and $III$, we again bound $p_k$ by a combination of Lemmas 2 and 21. We introduce the following function class, for $1 \leq k \leq s$,

$$\mathcal{F}_k := \{(\boldsymbol{x}, y) \mapsto \{I(y \leq \boldsymbol{x}^T\boldsymbol{\beta}) - I(y \leq \boldsymbol{x}^T\boldsymbol{\beta}_0)\}x_k : \ \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| \leq \Delta_t, \|\boldsymbol{\beta}\|_0 \leq Cs\}.$$

Then we have

$$p_k \leq \mathbb{P}\left(n^{-1/2}\|\mathbb{G}_n\|_{\mathcal{F}_k} > u\right).$$

A simple entropy computation shows

$$\log \sup_Q N(\mathcal{F}_k, \|\cdot\|_{Q,2}, \epsilon\|F\|_{Q,2}) \leq Cs\log(p/\epsilon), \ 0 < \forall\epsilon \leq 1,$$

where $\sup_Q$ is taken over all finitely discrete distributions and $F$ is an envelope function of $\mathcal{F}_k$; see the proof of Theorem 1 in Belloni et al. (2015b) for related arguments. Now, we proceed as in $t \geq 1$ by applying Lemmas 2 and 21 to show Conclude that with probability at least $1 - Csn^{-s}$,

$$II \vee III \leq C\sqrt{n^{-1}\Delta_t K_n^2 s\log n}.$$

The analysis of term $I$ remains the same which finishes the proof. □

We shall prove Lemma 22. We will focus on the case $t \geq 1$. The proof of $t = 0$ follows a similar argument and is skipped for brevity. Define $B_{\mathcal{S}}(l) := \{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| \leq l, \mathcal{S}_{\boldsymbol{\beta}} \subset \mathcal{S}\}$ . In the following, we condition on the event $\{\lambda_{t+1}\omega_{t+1,k} \geq \|(\nabla\tilde{\mathcal{L}}(\boldsymbol{\beta}_0; \hat{\boldsymbol{\beta}}_t)_{\mathcal{S}}\|_\infty \text{ for } 1 \leq k \leq s\}$ which has probability at least $1 - \delta - Csn^{-s} - \mathbb{P}(\|\hat{\boldsymbol{\beta}}_t - \boldsymbol{\beta}_0\| > \Delta_t \text{ or } \hat{\boldsymbol{\beta}}_t \neq \hat{\boldsymbol{\beta}}_t^\circ)$ by Lemma 28. We consider $\tilde{\mathcal{L}}_1(\boldsymbol{\beta}; \hat{\boldsymbol{\beta}}_t) - \tilde{\mathcal{L}}_1(\boldsymbol{\beta}_0; \hat{\boldsymbol{\beta}}_t)$ over $\partial B_{\mathcal{S}}(l)$. By Knight's identity Knight (1998),

$$\begin{aligned}
\tilde{\mathcal{L}}_1(\boldsymbol{\beta}; \hat{\boldsymbol{\beta}}_t) - \tilde{\mathcal{L}}_1(\boldsymbol{\beta}_0; \hat{\boldsymbol{\beta}}_t) &= \big\langle \nabla\tilde{\mathcal{L}}_1(\boldsymbol{\beta}_0; \hat{\boldsymbol{\beta}}_t), \boldsymbol{\beta} - \boldsymbol{\beta}_0 \big\rangle + \lambda_{t+1} \sum_{k=1}^{s} \omega_{t+1,k}(|\boldsymbol{\beta}_k| - |\boldsymbol{\beta}_{0k}|) \\
&\quad + \underbrace{\frac{1}{n}\sum_{j=1}^{n} \int_0^{\boldsymbol{X}_{1j}^T(\boldsymbol{\beta}-\boldsymbol{\beta}_0)} I\{y_{1j} - \boldsymbol{X}_{1j}^T\boldsymbol{\beta}_0 \leq u\} - I\{y_{1j} - \boldsymbol{X}_{1j}^T\boldsymbol{\beta}_0 \leq 0\} du}_{I} \\
&\geq \mathbb{E}[I] - C'\sqrt{s}(\sqrt{\eta_t s K_n^2 \log n/n} \vee \sqrt{K_n^2 \log n/N})l + (I - \mathbb{E}[I]) \\
&=: A_1 + A_2 + A_3.
\end{aligned}$$

Let $F(\cdot \mid \boldsymbol{x})$ denote the conditional distribution function of $\epsilon$ in the quantile regression model and $\delta = \boldsymbol{\beta} - \boldsymbol{\beta}_0$. If $l \leq \frac{3q\underline{f}\lambda_{\min}}{2\bar{f}'\lambda_{\max}^{3/2}}$, then

$$\begin{aligned}
A_1 &= \mathbb{E}\big[\int_0^{\boldsymbol{X}^T\delta} F(u \mid \boldsymbol{X}) - F(0 \mid \boldsymbol{X}) du\big] = \mathbb{E}\big[\int_0^{\boldsymbol{X}^T\delta} f(0 \mid \boldsymbol{X})u + \frac{1}{2}f'(\tilde{u} \mid \boldsymbol{X})u^2 du\big] \\
&\geq \frac{\underline{f}}{2}\mathbb{E}\big[(\boldsymbol{X}^T\delta)^2\big] - \frac{\bar{f}'}{6}\mathbb{E}\big[|\boldsymbol{X}^T\delta|^3\big] \geq \frac{\underline{f}\lambda_{\min}}{2}l^2 - \frac{\bar{f}'\lambda_{\max}^{3/2}}{6q}l^3 \\
&\geq \frac{\underline{f}\lambda_{\min}l^2}{4}.
\end{aligned}$$

For $A_3$, define

$$\mathcal{F} := \left\{(\boldsymbol{x}, y) \mapsto \boldsymbol{x}^T(\boldsymbol{\beta} - \boldsymbol{\beta}_0) \int_0^1 \big\{I\big(y \leq z\boldsymbol{x}^T\boldsymbol{\beta} + (1-z)\boldsymbol{x}^T\boldsymbol{\beta}_0\big) - I(y \leq \boldsymbol{x}^T\boldsymbol{\beta}_0)\big\} dz : \boldsymbol{\beta} \in \partial B_{\mathcal{S}}(l)\right\}.$$

Arguing similarly to the proof of Lemma 28 for term II, we have that with probability at least $1 - n^{-s}$,

$$A_3 \leq n^{-1/2}\|\mathbb{G}\|_{\mathcal{F}} \leq C''\sqrt{n^{-1}s^{3/2}K_n l^3 \log l^{-1}}.$$

Combining these bounds for $A_1$ and $A_3$, if we take

$$l = C\left(\sqrt{\frac{\eta_t s^2 K_n^2 \log n}{n}} \bigvee \sqrt{\frac{s K_n^2 \log(n/\delta)}{N}}\right)$$

for a sufficiently large constant $C$ (independent of $(n,t)$) such that

$$\frac{\underline{f}\lambda_{\min} l^2}{4} - C'\sqrt{s}(\sqrt{\eta_t s K_n^2 \log n/n} \vee \sqrt{K_n^2 \log n/N})l - C''\sqrt{n^{-1}s^{3/2}K_n l^3 \log l^{-1}} > 0,$$

then $l \leq \frac{3q\underline{f}\lambda_{\min}}{2\bar{f}'\lambda_{\max}^{3/2}}$ by assumption, and we have that with probability at least $1 - \delta - Csn^{-s} - \mathbb{P}(\|\hat{\boldsymbol{\beta}}_t - \boldsymbol{\beta}_0\| > \Delta_t \text{ or } \hat{\boldsymbol{\beta}}_t \neq \hat{\boldsymbol{\beta}}_t^\circ)$,

$$\inf_{\boldsymbol{\beta} \in \partial B_{\mathcal{S}}(l)} \tilde{\mathcal{L}}_1(\boldsymbol{\beta}; \hat{\boldsymbol{\beta}}_t) - \tilde{\mathcal{L}}_1(\boldsymbol{\beta}_0; \hat{\boldsymbol{\beta}}_t) > 0.$$

The desired conclusion follows from convexity of $\tilde{\mathcal{L}}(\boldsymbol{\beta}; \hat{\boldsymbol{\beta}}_t)$ in $\boldsymbol{\beta}$ and the definition of $\hat{\boldsymbol{\beta}}_{t+1}^\circ$. □

## Proof of Lemma 23

For $t \geq 0$, by KKT condition, $\hat{\boldsymbol{\beta}}_{t+1}^\circ$ is the global solution of (3.6) if for $s+1 \leq k \leq p$,

$$|[\nabla \tilde{\mathcal{L}}_1(\hat{\boldsymbol{\beta}}_{t+1}^\circ; \hat{\boldsymbol{\beta}}_t)]_k| - \lambda_{t+1}\omega_{t+1,k} < 0, \tag{C.10}$$

where $[\nabla \tilde{\mathcal{L}}_1(\hat{\boldsymbol{\beta}}_{t+1}^\circ; \hat{\boldsymbol{\beta}}_t)]_k$ denotes the $k$-th coordinate of $\nabla \tilde{\mathcal{L}}_1(\hat{\boldsymbol{\beta}}_{t+1}^\circ; \hat{\boldsymbol{\beta}}_t)$. Next we show $(C.10)$ holds with high probability.

We note

$$\begin{aligned}
|[\nabla \tilde{\mathcal{L}}_1(\hat{\boldsymbol{\beta}}_{t+1}^\circ; \hat{\boldsymbol{\beta}}_t)]_k| \leq & \left|\mathbb{E}[(\tau - I\{y \leq \boldsymbol{X}^T\beta\})X_k]\big|_{\beta=\hat{\boldsymbol{\beta}}_{t+1}^\circ}\right| + \left|\frac{1}{m}\sum_{i=1}^m [\nabla \mathcal{L}_i(\boldsymbol{\beta}_0)]_k\right| \\
& + \left|[\nabla \mathcal{L}_1(\hat{\boldsymbol{\beta}}_{t+1}^\circ) - \nabla \mathcal{L}_1(\boldsymbol{\beta}_0)]_k - \mathbb{E}[(\tau - I\{y \leq \boldsymbol{X}^T\beta\})X_k]\big|_{\beta=\hat{\boldsymbol{\beta}}_{t+1}^\circ}\right| \\
& + \left|[\nabla \mathcal{L}_1(\hat{\boldsymbol{\beta}}_t) - \nabla \mathcal{L}_1(\boldsymbol{\beta}_0)]_k - \mathbb{E}[(\tau - I\{y \leq \boldsymbol{X}^T\beta\})X_k]\big|_{\beta=\hat{\boldsymbol{\beta}}_t}\right| \\
& + \left|\frac{1}{m}\sum_{i=1}^m [(\nabla \mathcal{L}_i(\hat{\boldsymbol{\beta}}_t) - \nabla \mathcal{L}_i(\boldsymbol{\beta}_0))]_k - \mathbb{E}[(\tau - I\{y \leq \boldsymbol{X}^T\beta\})X_k]\big|_{\beta=\hat{\boldsymbol{\beta}}_t}\right| \\
& := A_1 + A_2 + \cdots + A_5.
\end{aligned}$$

For $A_1$,

$$\begin{aligned} A_1 &\le \bar{f}\mathbb{E}_{\boldsymbol{X}}\big[\big|\boldsymbol{X}_k\boldsymbol{X}^T(\hat{\boldsymbol{\beta}}^\circ_{t+1}-\boldsymbol{\beta}_0)\big|\big] \\ &\le \bar{f}\big(\mathbb{E}_{\boldsymbol{X}}[\{\boldsymbol{X}_k\boldsymbol{X}^T(\hat{\boldsymbol{\beta}}^\circ_{t+1}-\boldsymbol{\beta}_0)\}^2]\big)^{1/2} \\ &= O(a_n\|\hat{\boldsymbol{\beta}}^\circ_{t+1}-\boldsymbol{\beta}_0\|). \end{aligned}$$

For the rest of the terms, the analysis is similar to Lemma 28. Finally we can show, with probability at least $1-\delta-Cpn^{-s}-I\{t\ge 1\}\mathbb{P}(\|\hat{\boldsymbol{\beta}}_t-\boldsymbol{\beta}_0\|>\Delta_t \text{ or } \hat{\boldsymbol{\beta}}_t\neq\hat{\boldsymbol{\beta}}^\circ_t)-I\{t=0\}\mathbb{P}(\|\hat{\boldsymbol{\beta}}_0-\boldsymbol{\beta}_0\|>\Delta_0 \text{ or } \|\hat{\boldsymbol{\beta}}_0\|_0\ge Cs)$, for any $s+1\le k\le p$,

$$|[\nabla\tilde{\mathcal{L}}_1(\hat{\boldsymbol{\beta}}^\circ_{t+1};\hat{\boldsymbol{\beta}}_t)]_k| = O\big(K_n\sqrt{s\log p}\big((a_n\sqrt{s}\vee 1)\sqrt{\Delta_t/n}\vee N^{-1/2}\big)\big) < \lambda_{t+1}\omega_{t+1,k}.$$

In view of ($C$.10), this implies $\hat{\boldsymbol{\beta}}_{t+1}=\hat{\boldsymbol{\beta}}^\circ_{t+1}$.

## Proof of Lemma 24

The following $\ell^\infty$-bound of the gradient of the loss function will be useful.

**Lemma 29.** *Suppose Assumption 6 holds and fix any $t\ge 0$. Then with probability $1-\delta-\mathbb{P}(\|\hat{\boldsymbol{\beta}}_t-\boldsymbol{\beta}_0\|_1>\Delta_t)$, we have*

$$\|\nabla\tilde{\mathcal{L}}_1(\boldsymbol{\beta}_0;\hat{\boldsymbol{\beta}}_t)\|_\infty \le C\left[V_nK_n\sqrt{\frac{\log(p/\delta)}{N}}+K_n^2\sqrt{\frac{\log(p/\delta)}{n}}\cdot\Delta_t\right]. \tag{C.11}$$

*Proof.* First we have

$$\begin{aligned} &\|\nabla\tilde{\mathcal{L}}_1(\boldsymbol{\beta}_0,\hat{\boldsymbol{\beta}}_t)\|_\infty = \|\nabla\mathcal{L}_1(\boldsymbol{\beta}_0)-\nabla\mathcal{L}_1(\hat{\boldsymbol{\beta}}_t)+\frac{1}{m}\sum_{i=1}^m\nabla\mathcal{L}_i(\hat{\boldsymbol{\beta}}_t)\|_\infty \\ &\le \left\|\frac{1}{m}\sum_{i=1}^m\nabla\mathcal{L}_i(\boldsymbol{\beta}_0)\right\|_\infty + \left\|\nabla\mathcal{L}_1(\hat{\boldsymbol{\beta}}_t)-\nabla\mathcal{L}_1(\boldsymbol{\beta}_0)+\frac{1}{m}\sum_{i=1}^m(\nabla\mathcal{L}_i(\hat{\boldsymbol{\beta}}_t)-\nabla\mathcal{L}_i(\boldsymbol{\beta}_0))\right\|_\infty \\ &:= I+II \end{aligned}$$

First, we look at term II.

$$
\begin{aligned}
II &= 2\left\| n^{-1}\textstyle\sum_{j=1}^{n} \boldsymbol{X}_{1j}\boldsymbol{X}_{1j}^T(\hat{\boldsymbol{\beta}}_t - \boldsymbol{\beta}_0) + (mn)^{-1}\sum_{i=1}^{m}\sum_{j=1}^{n} \boldsymbol{X}_{ij}\boldsymbol{X}_{ij}^T(\hat{\boldsymbol{\beta}}_t - \boldsymbol{\beta}_0)\right\|_\infty \\
&\le 2\big(\|n^{-1}\textstyle\sum_{j=1}^{n} \boldsymbol{X}_{1j}\boldsymbol{X}_{1j}^T - \boldsymbol{\Sigma}\|_\infty + \|(mn)^{-1}\sum_{i=1}^{m}\sum_{j=1}^{n} \boldsymbol{X}_{ij}\boldsymbol{X}_{ij}^T - \boldsymbol{\Sigma}\|_\infty\big)\cdot\|\hat{\boldsymbol{\beta}}_t - \boldsymbol{\beta}_0\|_1 \\
&\le CK_n^2\Big(\sqrt{\frac{\log(p/\delta)}{n}} + \sqrt{\frac{\log(p/\delta)}{N}}\Big)\cdot\|\hat{\boldsymbol{\beta}}_t - \boldsymbol{\beta}_0\|_1, \text{ (with probability } 1-\delta/2),
\end{aligned}
$$

where the last inequality is due to Hoeffding's inequality, (cf. Theorem 2.8 in Boucheron et al. (2013)) and a union bound.

Now we consider term I.

$$
\begin{aligned}
\mathbb{P}\left(I > t\right) &\le p\cdot\max_{1\le k\le p}\mathbb{P}\left(\Big|\frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} 2(y_{ij} - \boldsymbol{X}_{ij}^T\boldsymbol{\beta}_0)X_{ijk}\Big| > t\right) \\
&\le p\cdot 2\exp\Big\{-\frac{Nt^2}{2V_n^2K_n^2}\Big\},
\end{aligned}
$$

where we used Hoeffding's inequality with the observation $|y - \boldsymbol{X}^T\boldsymbol{\beta}_0|\cdot|X_k| \le V_nK_n$. Hence, with probability $1-\delta/2$,

$$
I \le CV_nK_n\sqrt{\frac{\log(p/\delta)}{N}}
$$

Combining the analysis of terms I and II, we finally prove the lemma. □

Define cone $\mathcal{C} := \{\boldsymbol{\theta}\in\mathbb{R}^p : \ \|\boldsymbol{\theta}_{\mathcal{S}^c}\|_1 \le 3\|\boldsymbol{\theta}_{\mathcal{S}}\|_1\}$. We will also need the following cone condition of the solution.

**Lemma 30** (Cone condition)**.** *Suppose the conditions of Lemma 29 hold. If we take*

$$
\lambda_{t+1} \ge 2C\left[V_nK_n\sqrt{\frac{\log(p/\delta)}{N}} + K_n^2\sqrt{\frac{\log(p/\delta)}{n}}\cdot\Delta_t\right],
$$

*where the constant $C$ is the same as in Lemma 29. Then we have $\hat{\boldsymbol{\beta}}_{t+1} - \boldsymbol{\beta}_0 \in \mathcal{C}$ with probability at least $1-\delta-\mathbb{P}(\|\hat{\boldsymbol{\beta}}_t - \boldsymbol{\beta}_0\|_1 > \Delta_t)$.*

*Proof.* By the optimality of $\hat{\boldsymbol{\beta}}_t$,

$$\tilde{\mathcal{L}}_1(\hat{\boldsymbol{\beta}}_{t+1};\hat{\boldsymbol{\beta}}_t)+\lambda_{t+1}\|\hat{\boldsymbol{\beta}}_{t+1}\|_1-\tilde{\mathcal{L}}_1(\boldsymbol{\beta}_0;\hat{\boldsymbol{\beta}}_t)-\lambda_{t+1}\|\boldsymbol{\beta}_0\|_1\le 0,$$

which further implies

$$\begin{aligned}
0 &\ge \nabla\tilde{\mathcal{L}}_1(\boldsymbol{\beta}_0;\hat{\boldsymbol{\beta}}_t)^T(\hat{\boldsymbol{\beta}}_{t+1}-\boldsymbol{\beta}_0)+\lambda_{t+1}(\|\hat{\boldsymbol{\beta}}_{t+1}\|_1-\|\boldsymbol{\beta}_0\|_1)\\
&\ge -\|\nabla\tilde{\mathcal{L}}_1(\boldsymbol{\beta}_0;\hat{\boldsymbol{\beta}}_t)\|_\infty\,\|\hat{\boldsymbol{\beta}}_{t+1}-\boldsymbol{\beta}_0\|_1+\lambda_{t+1}(\|\hat{\boldsymbol{\beta}}_{t+1}\|_1-\|\boldsymbol{\beta}_0\|_1)\\
&\ge -\lambda_{t+1}/2\cdot\|\hat{\boldsymbol{\beta}}_{t+1}-\boldsymbol{\beta}_0\|_1+\lambda_{t+1}(\|\hat{\boldsymbol{\beta}}_{t+1}\|_1-\|\boldsymbol{\beta}_0\|_1).\quad \text{(w.p. } 1-\delta-\mathbb{P}(\|\hat{\boldsymbol{\beta}}_t-\boldsymbol{\beta}_0\|_1>\Delta_t))
\end{aligned}$$

This combined with the observation $\|\hat{\boldsymbol{\beta}}_{t+1}\|_1-\|\boldsymbol{\beta}_0\|_1\ge\|(\hat{\boldsymbol{\beta}}_{t+1}-\boldsymbol{\beta}_0)_{\mathcal{S}^c}\|_1-\|(\hat{\boldsymbol{\beta}}_{t+1}-\boldsymbol{\beta}_0)_{\mathcal{S}}\|_1$ yields the conclusion. □

Now we are ready to prove Lemma 24. Define $B_l:=\{\boldsymbol{\beta}:\boldsymbol{\beta}-\boldsymbol{\beta}_0\in\mathcal{C},\ \|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|\le l\}$. In the following, we conditional on the event $E:=\{\|\hat{\boldsymbol{\beta}}_t-\boldsymbol{\beta}_0\|_1\le\Delta_t,\lambda_{t+1}\ge\|\nabla\tilde{\mathcal{L}}_1(\boldsymbol{\beta}_0;\hat{\boldsymbol{\beta}}_t)\|_\infty\}$ which has probability $1-\delta-\mathbb{P}(\|\hat{\boldsymbol{\beta}}_t-\boldsymbol{\beta}_0\|_1>\Delta_t)$. We consider $\tilde{\mathcal{L}}_1(\boldsymbol{\beta};\hat{\boldsymbol{\beta}}_t)-\tilde{\mathcal{L}}_1(\boldsymbol{\beta}_0;\hat{\boldsymbol{\beta}}_t)$ over $\partial B_l$.

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\beta};\hat{\boldsymbol{\beta}}_t)-\mathcal{L}(\boldsymbol{\beta}_0;\hat{\boldsymbol{\beta}}_t) &= \big\langle\nabla\tilde{\mathcal{L}}(\boldsymbol{\beta}_0;\hat{\boldsymbol{\beta}}_t),\boldsymbol{\beta}-\boldsymbol{\beta}_0\big\rangle+\lambda_{t+1}(\|\boldsymbol{\beta}\|_1-\|\boldsymbol{\beta}_0\|_1)\\
&\quad+(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^T(n^{-1}\textstyle\sum_{j=1}^n\boldsymbol{X}_{1j}\boldsymbol{X}_{1j}^T)(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\\
&\ge -\big(\|\nabla\tilde{\mathcal{L}}(\boldsymbol{\beta}_0;\hat{\boldsymbol{\beta}}_t)\|_\infty+\lambda_{t+1}\big)\cdot\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|_1+C'\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|^2\quad\text{(w.p. } 1-\nu_n)\\
&\ge C'l^2-C\lambda_{t+1}\sqrt{s}l\quad\text{(w.p. } 1-\delta-\nu_n-\mathbb{P}(\|\hat{\boldsymbol{\beta}}_t-\boldsymbol{\beta}_0\|_1>\Delta_t)),
\end{aligned}$$

where we use the observation for the sparse eigenvalue in the discussion after Assumption 6 in the first inequality. By choosing sufficiently large $l=O(\sqrt{s}\lambda_{t+1})$, we have, with probability $1-\delta-\nu_n-\mathbb{P}(\|\hat{\boldsymbol{\beta}}_t-\boldsymbol{\beta}_0\|_1>\Delta_t)$,

$$\inf_{\boldsymbol{\beta}\in\partial B_l}\mathcal{L}(\boldsymbol{\beta};\hat{\boldsymbol{\beta}}_t)-\mathcal{L}(\boldsymbol{\beta}_0;\hat{\boldsymbol{\beta}}_t)\ge 0.$$

By the convexity of the optimization problem and the definition of $\hat{\boldsymbol{\beta}}_{t+1}$, we have with probability $1-\delta-o(1)-\mathbb{P}(\|\hat{\boldsymbol{\beta}}_t-\boldsymbol{\beta}_0\|_1>\Delta_t)$,

$$\|\hat{\boldsymbol{\beta}}_{t+1}-\boldsymbol{\beta}_0\|_2\le C\Big[V_nK_n\sqrt{\frac{s\log(p/\delta)}{N}}+K_n^2\sqrt{\frac{s\log(p/\delta)}{n}}\cdot\Delta_t\Big].$$

We also note $\|\hat{\boldsymbol{\beta}}_{t+1} - \boldsymbol{\beta}_0\|_1 \leq C\sqrt{s}\|\hat{\boldsymbol{\beta}}_{t+1} - \boldsymbol{\beta}_0\|_2$ since $\hat{\boldsymbol{\beta}}_{t+1} - \boldsymbol{\beta}_0 \in \mathcal{C}$ under event $E$. This observation implies with probability $1 - \delta - \nu_n - \mathbb{P}(\|\hat{\boldsymbol{\beta}}_t - \boldsymbol{\beta}_0\|_1 > \Delta_t)$,

$$\|\hat{\boldsymbol{\beta}}_{t+1} - \boldsymbol{\beta}_0\|_1 \leq C\Big[V_n K_n s\sqrt{\frac{\log(p/\delta)}{N}} + sK_n^2\sqrt{\frac{\log(p/\delta)}{n}} \cdot \Delta_t\Big],$$

which finishes the proof. □

## Proof of Lemma 25

To prove the sparsity bound, we will need the following empirical prediction norm bound of $\hat{\boldsymbol{\beta}}_{t+1}$. We denote the design matrix of machine 1 as $\boldsymbol{D} := (\boldsymbol{X}_{11}, \ldots, \boldsymbol{X}_{1n})^T$.

**Corollary 4.** *Under the conditions of Theorem 7, with probability* $1 - p^{-1} - \gamma_n - \nu_n$ *independent of iteration* $t$*, we have*

$$n^{-1/2}\|D(\hat{\boldsymbol{\beta}}_{t+1} - \boldsymbol{\beta}_0)\| \leq C\sqrt{s}\lambda_{t+1}.$$

*Proof.* By the optimality of $\hat{\boldsymbol{\beta}}_{t+1}$,

$$\tilde{\mathcal{L}}_1(\hat{\boldsymbol{\beta}}_{t+1}; \hat{\boldsymbol{\beta}}_t) - \tilde{\mathcal{L}}_1(\boldsymbol{\beta}_0; \hat{\boldsymbol{\beta}}_t) \leq \lambda_{t+1}(\|\boldsymbol{\beta}_0\|_1 - \|\hat{\boldsymbol{\beta}}_{t+1}\|_1).$$

After simple rearranging of the above inequality, we have

$$\begin{aligned}
\frac{1}{n}\|\boldsymbol{D}(\hat{\boldsymbol{\beta}}_{t+1} - \boldsymbol{\beta}_0)\|^2 &\leq \Big|\big\langle \nabla\tilde{\mathcal{L}}_1(\boldsymbol{\beta}_0; \hat{\boldsymbol{\beta}}_t), \boldsymbol{\beta} - \boldsymbol{\beta}_0\big\rangle\Big| + \lambda_{t+1}(\|\boldsymbol{\beta}_0\|_1 - \|\hat{\boldsymbol{\beta}}_{t+1}\|_1) \\
&\leq C\lambda_{t+1}\|\hat{\boldsymbol{\beta}}_{t+1} - \boldsymbol{\beta}_0\|_1 \\
&\leq Cs\lambda_{t+1}^2,
\end{aligned}$$

with probability at least $1 - p^{-1} - \gamma_n - \nu_n$ by step 1 of the proof of Theorem 7. Taking square root on both sides gives the result. □

Now we prove Lemma 25. By KKT condition, for $j \in \hat{\mathcal{S}}_{t+1} \backslash \mathcal{S} := \Delta\hat{\mathcal{S}}_{t+1}$,

$$\Big( -\nabla \mathcal{L}_1(\hat{\boldsymbol{\beta}}_{t+1}) - \big(m^{-1}\textstyle\sum_{i=1}^m \nabla \mathcal{L}_i(\hat{\boldsymbol{\beta}}_t) - \nabla \mathcal{L}_1(\hat{\boldsymbol{\beta}}_t)\big)\Big)_j = \lambda_{t+1}\text{sign}(\hat{\boldsymbol{\beta}}_{t+1,j}).$$

This implies

$$\begin{aligned}
\sqrt{\hat{m}_{t+1}}\lambda_{t+1} &= \Big\| \Big( -\nabla \mathcal{L}_1(\hat{\boldsymbol{\beta}}_{t+1}) - \big(m^{-1}\textstyle\sum_{i=1}^m \nabla \mathcal{L}_i(\hat{\boldsymbol{\beta}}_t) - \nabla \mathcal{L}_1(\hat{\boldsymbol{\beta}}_t)\big)\Big)_{\Delta\hat{\mathcal{S}}_{t+1}} \Big\| \\
&\le \big\| \big(\nabla \tilde{\mathcal{L}}_1(\boldsymbol{\beta}_0; \hat{\boldsymbol{\beta}}_t)\big)_{\Delta\hat{\mathcal{S}}_{t+1}} \big\| + \big\| \big(\nabla \mathcal{L}_1(\hat{\boldsymbol{\beta}}_{t+1}) - \nabla \mathcal{L}_1(\boldsymbol{\beta}_0)\big)_{\Delta\hat{\mathcal{S}}_{t+1}} \big\| \\
&\le \sqrt{\hat{m}_{t+1}} \|\nabla \tilde{\mathcal{L}}_1(\boldsymbol{\beta}_0; \hat{\boldsymbol{\beta}}_t)\|_\infty + 2 \big\| \big(n^{-1} D^T D(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\big)_{\Delta\hat{\mathcal{S}}_{t+1}} \big\| \\
&:= I + 2 \cdot II.
\end{aligned} \tag{C.12}$$

By properly choosing $\lambda_{t+1}$, we have with the same probability as in the step 1 of the proof of Theorem 7, $1 - p^{-1} - \gamma_n - \nu_n$, $I \le \sqrt{\hat{m}_{t+1}}\lambda_{t+1}/2$.

For II, denote $\mathcal{S}(\boldsymbol{\delta})$ the support set of a vector $\boldsymbol{\delta}$,

$$\begin{aligned}
II &= \sup_{\substack{\mathcal{S}(\boldsymbol{\delta}) = \Delta\hat{\mathcal{S}}_{t+1} \\ \|\boldsymbol{\delta}\| = 1}} n^{-1} \big| \boldsymbol{\delta}^T \boldsymbol{D}^T \boldsymbol{D}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \big| \\
&\le \sup_{\substack{\mathcal{S}(\boldsymbol{\delta}) = \Delta\hat{\mathcal{S}}_{t+1} \\ \|\boldsymbol{\delta}\| = 1}} n^{-1} \|\boldsymbol{D}\boldsymbol{\delta}\| \cdot \|\boldsymbol{D}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\| \\
&\le \bar{\phi}_{\max}(\hat{m}_{t+1}, \boldsymbol{\Sigma}_n)^{1/2} \cdot C\sqrt{s}\lambda_{t+1} \quad \text{with probability } 1 - p^{-1} - \gamma_n - \nu_n,
\end{aligned}$$

where the last inequality is due to Corollary 4. Combining the above analysis of I and II, we prove the result by a simple rearrangement in (C.12).

## Proof of Lemma 26

We proceed as the Step 1 of the proof of Theorem 2 in Belloni et al. (2015b). We note the conditions required there are satisfied under our conditions and the conclusion easily follows by keeping track of the convergence rate of the residue terms in the proof.

### Proof of Lemma 27

We define $\tilde{f}_\epsilon(0) := N^{-1}\sum_{i=1}^N h^{-1}K(\epsilon_i/h)$ and we have

$$|\hat{f}_\epsilon(0) - f_\epsilon(0)| \le |\hat{f}_\epsilon(0) - \tilde{f}_\epsilon(0)| + |\tilde{f}_\epsilon(0) - f_\epsilon(0)| := I + II$$

By a conventional analysis from the kernel density estimator (cf. Wasserman (2006)), we can show

$$II = O_P\big(h^2 + (Nh)^{-1/2}\big).$$

For I,

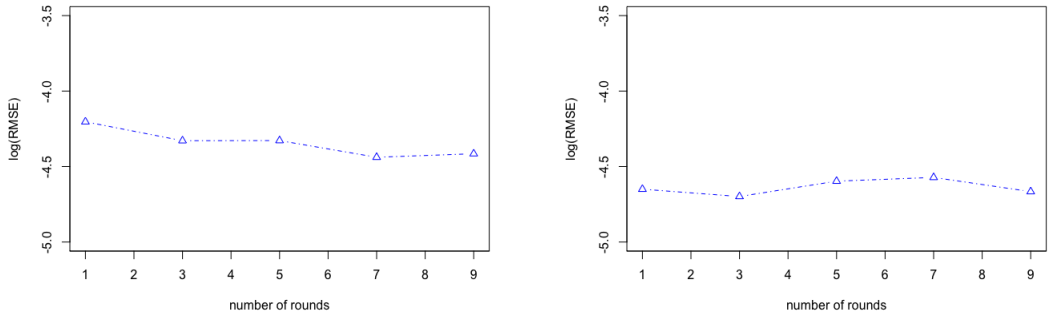
$$\begin{aligned}
I &= \Big|N^{-1}\textstyle\sum_{j=1}^N h^{-1}\big(K(\hat{\epsilon}_j/h) - K(\epsilon_i/h)\big)\Big| \\
&= |\textstyle\sum_{j=1}^N K_j' \cdot N^{-1}h^{-2}(\hat{\epsilon}_j - \epsilon_j)| \quad \text{for some } K_j' \text{ between } K'(\hat{\epsilon}_j/h) \text{ and } K(\epsilon_j/h) \\
&\le \sqrt{\textstyle\sum_{j=1}^N (K_j')^2} \cdot N^{-1}h^{-2}\sqrt{\textstyle\sum_{j=1}^N \big(\boldsymbol{X}_j^T(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}})\big)^2} \\
&\le CN^{-1/2}h^{-2}\|\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}\| \quad \text{with probability } 1 - o(1) \\
&= O_P\big(N^{-1/2}h^{-2}\rho_N\big).
\end{aligned}$$

Finally we combine the analysis above to finish the proof.

## C.6 Sensitivity of the number of rounds

In this section, we study the performance of the multi-round estimator using different number of rounds. We generate the data as in Section 3.4.2 with $\epsilon \overset{i.i.d.}{\sim} \exp(1)$ (the results of the other two scenarios are similar). We fix the total sample size $N = 12000$ which is the maximum considered in our simulation study. Two local sample sizes $n = 100$ and $n = 300$ are considered which correspond to $m = 120$ and $m = 40$, respectively. We repeat the simulation for 400

times for each different number of rounds increasing from $1$ to $9$. We plot the root mean squared error (RMSE) versus the number of rounds and include the performance of the centralized quantile regression estimator as a benchmark. The plots are shown in Figure C.1. From the plots, the multi-round estimator becomes stable and achieves desirable convergence rate in relatively few rounds for the two considered local sample sizes.




(a) $n = 100$ (b) $n = 300$

Figure C.1: RMSE plots of the proposed multi-round debiased estimator using different number of rounds with $N = 12000$.

## C.7 Comparison of the computational time

In this section, we compare the computational time of the one-shot estimator and the multi-round estimator. We generate the data as in Section 3.4.2 with $\epsilon \overset{i.i.d.}{\sim} \exp(1)$. We fix the local sample size $n = 300$ and dimension $p = 400$. We increase the number of machines ($m$) from $10$ to $50$. We compare the average computational time of the one-shot estimator and the multi-round estimator using iteration round from $1$ to $5$ based on $400$ repetitions. For each repetition, we record the total running time of tunning parameter selection and computing the

estimator. We also provide the corresponding RMSE comparison. The results are presented in Figure C.2.

From the plots, we can see the computational time of the multi-round estimator is insensitive to the increase of the number of machines when the number of rounds is fixed while the computational time of the one-shot estimator almost increases linearly with the number of machines. Particularly, the one-shot estimator requires significantly more computational time than the multi-round estimator when the number of machines is large. It is also interesting to note the computational time of the multi-round estimator even decreases when the number of the machines increase from $30$ to $50$ and this is more evident for estimators with more rounds. After a more careful research, we find that the decrease of the running time is mainly due to less running time to select the tunning parameters for the multi-round lasso estimator while the computational time for computing the estimators only slightly increases. We conjecture that the decrease of the tunning parameter selection time for the lasso is because of a faster convergence of the accelerated proximal gradient algorithm at certain grid points of the tunning parameter. Meanwhile, the multi-round estimator results in smaller RMSE than the one-shot estimator uniformly in all the scenarios which supports our theory. The performance of the multi-round estimator using different rounds is similar to each other and this phenomenon is more evident when the number of machines is large. This again suggests the multi-round estimator becomes stable and achieves desirable convergence rate in relatively few rounds even for a large number of machines. Combining the comparison of the computational and statistical performance above, we can see that the multi-round estimator achieves a better estimation performance than the one-shot estimator while requiring less computational time.

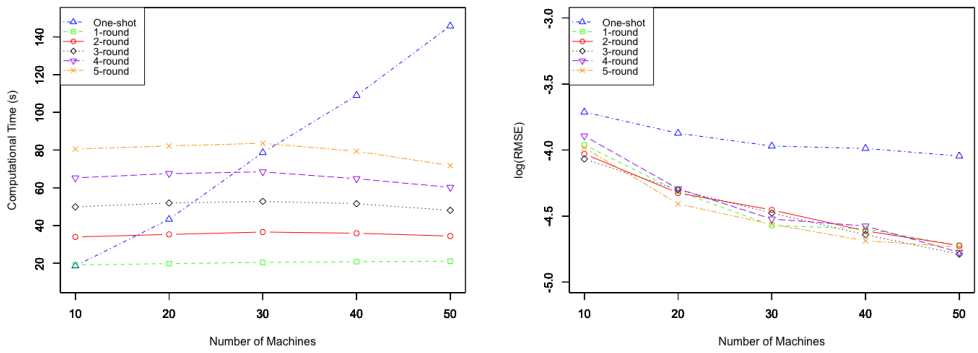



Figure C.2: Left: computational time comparison between the one-shot estimator and the multi-round estimator; Right: RMSE comparison between the one-shot estimator and the multi-round estimator.

## C.8 Formulating the multi-round quantile regression as linear programming problems

In this section we show that optimization problem at the $t$-th round of the multi-round quantile regression can be formulated as a linear programming. First, we recall that at the $t$-th round we solve the following optimization problem, for a given $\hat{\boldsymbol{\beta}}_t$,

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{t+1} &= \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p}\left\{\mathcal{L}_1(\boldsymbol{\beta}) + \left\langle \frac{1}{m}\sum_{i=1}^{m}\nabla\mathcal{L}_i(\hat{\boldsymbol{\beta}}_t) - \nabla\mathcal{L}_1(\hat{\boldsymbol{\beta}}_t), \boldsymbol{\beta}\right\rangle + \lambda_{t+1}\sum_{k=1}^{p}\omega_{t+1,k}|\beta_k|\right\} \\ &:= \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p}\left\{\mathcal{L}_1(\boldsymbol{\beta}) + \langle D_t, \boldsymbol{\beta}\rangle + \lambda_{t+1}\sum_{k=1}^{p}\omega_{t+1,k}|\beta_k|\right\},\end{aligned}$$

where $\mathcal{L}_i(\boldsymbol{\beta}) := n^{-1}\sum_{j=1}^{n}\ell(\boldsymbol{X}_{ij}, y_{ij}, \boldsymbol{\beta})$, $\ell(\boldsymbol{X}, y, \boldsymbol{\beta}) = \{\tau - I(y \le \boldsymbol{X}^T\boldsymbol{\beta}\})(y - \boldsymbol{X}^T\boldsymbol{\beta})$ and $\nabla\mathcal{L}_i(\boldsymbol{\beta}) = n^{-1}\sum_{j=1}^{n}\{I(y_{ij} \le \boldsymbol{X}_{ij}^T\boldsymbol{\beta}) - \tau\}\boldsymbol{X}_{ij}$.

We define the following augmented dataset

$$(y_j^*, \boldsymbol{X}_j^*) = \begin{cases} (y_{1j}, \boldsymbol{X}_{1j}) & 1 \le j \le n \\ (0, n\lambda_{t+1}\omega_{t+1,j-n}\boldsymbol{e}_{j-n}) & n+1 \le j \le n+p \\ (0, -n\lambda_{t+1}\omega_{t+1,j-n-p}\boldsymbol{e}_{j-n-p}) & n+p+1 \le j \le n+2p \end{cases},$$

where $\boldsymbol{e}_k$ is the $k$-th canonical basis vector of $p$-dimensional Euclidean space. Hence, we have

$$\hat{\boldsymbol{\beta}}_{t+1} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \left\{ n^{-1} \sum_{j=1}^{n+2p} \ell(\boldsymbol{X}_j^*, y_j^*, \boldsymbol{\beta}) + \langle D_t, \boldsymbol{\beta} \rangle \right\}. \tag{C.13}$$

Now we formulate the above optimization as a linear programming. Recall that the standard form of LP is as follows,

$$\begin{aligned} \underset{\boldsymbol{z}}{\text{minimize}} \qquad & \boldsymbol{c}^T \boldsymbol{z} \\ \text{subject to} \qquad & \boldsymbol{A}\boldsymbol{z} = \boldsymbol{b} \quad \text{and} \quad \boldsymbol{z} \geq 0. \end{aligned}$$

We take

$$\begin{aligned} \boldsymbol{c} &= [D_t, -D_t, (\tau/n)\mathbf{1}_{n+2p}, (\tau/n)\mathbf{1}_{n+2p}]^T, \\ \boldsymbol{A} &= [\boldsymbol{X}^*, -\boldsymbol{X}^*, \boldsymbol{I}_{n+2p}, -\boldsymbol{I}_{n+2p}], \quad \boldsymbol{b} = \boldsymbol{y}^*, \end{aligned}$$

where $\mathbf{1}_{n+2p}$ is the $n+2p$ dimensional vector of 1, $\boldsymbol{I}_{n+2p}$ is the $(n+2p)$ by $(n+2p)$ identity matrix, $y^*$ is the column vector collecting all $\boldsymbol{y}_j^*$ and $\boldsymbol{X}^*$ is the augmented design matrix with $\boldsymbol{X}_j^{*T}$ as rows. Then it is not difficult to see the solution of the LP problem satisfies

$$\boldsymbol{z} = [\hat{\boldsymbol{\beta}}_{t+1}^+, \hat{\boldsymbol{\beta}}_{t+1}^-, \boldsymbol{u}, \boldsymbol{v}],$$

where $\hat{\boldsymbol{\beta}}_{t+1}^+$ and $\hat{\boldsymbol{\beta}}_{t+1}^-$ are positive and negative parts of $\hat{\boldsymbol{\beta}}_{t+1}$ respectively; $\boldsymbol{u}$ and $\boldsymbol{v}$ are positive and negative parts of $\left(\boldsymbol{y}^* - \boldsymbol{X}^* \hat{\boldsymbol{\beta}}_{t+1}\right)$ respectively.

## BIBLIOGRAPHY


Adamczak, R. (2008). A tail inequality for suprema of unbounded empirical processes with applications to markov chains. *Electronic Journal of Probability*, 13:1000–1034.

Ai, M., Yu, J., Zhang, H., and Wang, H. (2018). Optimal subsampling algorithms for big data regressions. *arXiv preprint arXiv:1806.06761*.

Aitkin, M. A., Aitkin, M., Francis, B., and Hinde, J. (2005). *Statistical modelling in GLIM 4*, volume 32. OUP Oxford.

Aliprantis, C. D. and Border, K. C. (2006). *Infinite Dimensional Analysis: a Hitchhiker's Guide*. Springer, Berlin; London.

Angrist, J., Chernozhukov, V., and Fernández-Val, I. (2006). Quantile regression under misspecification, with an application to the us wage structure. *Econometrica*, 74(2):539–563.

ApS, M. (2021). *MOSEK Rmosek package 9.2.40*.

Autor, D. H., Katz, L. F., and Kearney, M. S. (2008). Trends in us wage inequality: Revising the revisionists. *The Review of Economics and Statistics*, 90(2):300–323.

Bamford, S. P., Rojas, A. L., Nichol, R. C., Miller, C. J., Wasserman, L., Genovese, C. R., and Freeman, P. E. (2008). Revealing components of the galaxy population through non-parametric techniques. *Monthly Notices of the Royal Astronomical Society*, 391(2):607–616.

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

Banerji, M., Lahav, O., Lintott, C. J., Abdalla, F. B., Schawinski, K., Bamford, S. P., Andreescu, D., Murray, P., Raddick, M. J., and Slosar, A. (2010). Galaxy zoo: reproducing galaxy morphologies via machine learning. *Monthly Notices of the Royal Astronomical Society*, 406(1):342–353.

Battey, H., Tan, K. M., and Zhou, W.-X. (2021). Communication-efficient distributed quantile regression with optimal statistical guarantees. *arXiv preprint arXiv:2110.13113*.

Belloni, A., Chen, D., Chernozhukov, V., and Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80(6):2369–2429.

Belloni, A. and Chernozhukov, V. (2011). $\ell_1$-penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics*, 39(1):82–130.

Belloni, A., Chernozhukov, V., Chetverikov, D., and Fernndez-Val, I. (2019a). Conditional quantile processes based on series or many regressors. *Journal of Econometrics*, 213(1):4 – 29.

Belloni, A., Chernozhukov, V., Chetverikov, D., and Kato, K. (2015a). Some new asymptotic theory for least squares series: Pointwise and uniform results. *Journal of Econometrics*, 186:345–366.

Belloni, A., Chernozhukov, V., and Kato, K. (2015b). Uniform post-selection inference for least absolute deviation regression and other z-estimation problems. *Biometrika*, 102(1):77–94.

Belloni, A., Chernozhukov, V., and Kato, K. (2019b). Valid post-selection inference in high-dimensional approximately sparse quantile regression models. *Journal of the American Statistical Association*, 114(526):749–758.

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732.

Bissantz, N., Dümbgen, L., Holzmann, H., and Munk, A. (2007). Non-parametric confidence bands in deconvolution density estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(3):483–506.

Bondell, H. D., Reich, B. J., and Wang, H. (2010). Noncrossing quantile regression curve estimation. *Biometrika*, 97(4):825–838.

Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press.

Bound, J. and Krueger, A. B. (1991). The extent of measurement error in longitudinal earnings data: Do two wrongs make a right? *Journal of Labor Economics*, 9(1):1–24.

Bousquet, O. (2002). A bennett concentration inequality and its application to suprema of empirical processes. *Comptes Rendus Mathematique*, 334(6):495–500.

Buchinsky, M. (1994). Changes in the u.s. wage structure 1963-1987: Application of quantile regression. *Econometrica*, 62(2):405–458.

Bühlmann, P. and Van De Geer, S. (2011). *Statistics for High-dimensional Data: Methods, Theory and Applications*. Springer Science & Business Media.

Cai, T. T. and Guo, Z. (2018). Semi-supervised inference for explained variance in high-dimensional linear regression and its applications. *arXiv preprint arXiv:1806.06179*.

Chacón, J. (2018). The modal age of statistics. arXiv:1807.02789.

Chakrabortty, A. and Cai, T. (2018). Efficient and adaptive linear regression in semi-supervised settings. *The Annals of Statistics*, 46(4):1541–1572.

Chapelle, O., Schlkopf, B., and Zien, A. (2010). *Semi-Supervised Learning*. The MIT Press, 1st edition.

Chen, X. and Kato, K. (2020). Jackknife multiplier bootstrap: finite sample approximations to the u-process supremum with applications. *Probability Theory and Related Fields*, 176(3):1097–1163.

Chen, X., Lee, J. D., Li, H., and Yang, Y. (2021). Distributed estimation for principal component analysis: an enlarged eigenspace analysis. *Journal of the American Statistical Association*, pages 1–31.

Chen, X., Liu, W., Mao, X., and Yang, Z. (2020). Distributed high-dimensional regression under a quantile loss function. *Journal of Machine Learning Research*, 21(182):1–43.

Chen, X., Liu, W., and Zhang, Y. (2019). Quantile regression under memory constraint. *Annals of Statistics*, 47(6):3244–3273.

Chen, Y.-C. (2018). Modal regression using kernel density estimation: A review. *Wiley Interdisciplinary Reviews: Computational Statistics*, 10(4):e1431.

Chen, Y.-C., Genovese, C. R., Tibshirani, R. J., and Wasserman, L. (2016). Nonparametric modal regression. *The Annals of Statistics*, 44(2):489–514.

Cheng, Y. (1995). Mean shift, mode seeking, and clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(8):790–799.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey,

W., and Robins, J. (2016a). Double/debiased machine learning for treatment and causal parameters. *arXiv preprint arXiv:1608.00060*.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2014). Gaussian approximation of suprema of empirical processes. *The Annals of Statistics*, 42(4):1564–1597.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2016b). Empirical and multiplier bootstraps for suprema of empirical processes of increasing complexity, and related gaussian couplings. *Stochastic Processes and their Applications*, 126(12):3632–3651.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2017a). Central limit theorems and bootstrap in high dimensions. *The Annals of Probability*, 45(4):2309–2352.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2017b). Detailed proof of nazarov's inequality. arXiv:1711.10696.

Chernozhukov, V., Fernández-Val, I., and Galichon, A. (2010). Quantile and probability curves without crossing. *Econometrica*, 78(3):1093–1125.

Chernozhukov, V., Hansen, C., and Jansson, M. (2009). Finite sample inference for quantile regression models. *Journal of Econometrics*, 152:93–103.

Chung, K. (2001). *A Course in Probability Theory*. Elsevier Science.

Davidson, J. (1994). *Stochastic Limit Theory: An Introduction for Econometricians*. OUP Oxford.

Deng, H. and Zhang, C.-H. (2017). Beyond gaussian approximation: Bootstrap for maxima of sums of independent random vectors. *arXiv preprint arXiv:1705.09528*.

Drineas, P., Magdon-Ismail, M., Mahoney, M. W., and Woodruff, D. P. (2012). Fast approximation of matrix coherence and statistical leverage. *Journal of Machine Learning Research*, 13(1):3475–3506.

Drineas, P. and Mahoney, M. W. (2016). Randnla: randomized numerical linear algebra. *Communications of the ACM*, 59(6):80–90.

Drineas, P., Mahoney, M. W., and Muthukrishnan, S. (2006). Sampling algorithms for l2 regression and applications. In *Proceedings of the Seventeenth Annual ACM-SIAM Symposium on Discrete Algorithm*, SODA '06, pages 1127–1136, Philadelphia, PA, USA. Society for Industrial and Applied Mathematics.

Drineas, P., Mahoney, M. W., Muthukrishnan, S., and Sarlós, T. (2011). Faster least squares approximation. *Numerische Mathematik*, 117(2):219–249.

Einbeck, J. and Tutz, G. (2006). Modelling beyond regression functions: an application of multimodal regression to speed–flow data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 55(4):461–475.

Fan, J., Fan, Y., and Barut, E. (2014). Adaptive robust variable selection. *Annals of Statistics*, 42(1):324.

Fan, J., Wang, D., Wang, K., and Zhu, Z. (2019). Distributed estimation of principal eigenspaces. *Annals of Statistics*, 47(6).

Feng, Y., Fan, J., and Suykens, J. A. (2020). A statistical learning approach to modal regression. *Journal of Machine Learning Research*, 21(2):1–35.

Gutenbrunner, C. and Jurecková, J. (1992). Regression rank scores and regression quantiles. *The Annals of Statistics*, pages 305–330.

Hall, P. (1991). On convergence rates of suprema. *Probability Theory and Related Fields*, 89(4):447–455.

Hall, P. and Heyde, C. C. (2014). *Martingale limit theory and its application*. Academic press.

Hamidieh, K. (2018). A data-driven statistical model for predicting the critical temperature of a superconductor. *Computational Materials Science*, 154:346 – 354.

Han, Q. and Wellner, J. A. (2019). Convergence rates of least squares regression estimators with heavy-tailed errors. *Annals of Statistics*, 47(4):2286–2319.

Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical learning with sparsity: the lasso and generalizations*. CRC press.

He, X. (1997). Quantile curves without crossing. *The American Statistician*, 51(2):186–192.

He, X. (2017). Resampling methods. In *Handbook of quantile regression*, pages 7–19. Chapman and Hall/CRC.

He, X. and Shao, Q.-M. (1996). A general bahadur representation of m-estimators and its application to linear regression with nonstochastic designs. *The Annals of Statistics*, 24(6):2608–2630.

Heckman, D. S., Geiser, D. M., Eidell, B. R., Stauffer, R. L., Kardos, N. L., and Hedges, S. B. (2001). Molecular evidence for the early colonization of land by fungi and plants. *Science*, 293(5532):1129–1133.

Hedges, S. B. and Shah, P. (2003). Comparison of mode estimation methods and application in molecular clock analysis. *BMC Bioinformatics*, 4(1):31.

Ho, C., Damien, P., and Walker, S. (2017). Bayesian mode regression using mixtures of triangular densities. *Journal of Econometrics*, 197(2):273–283.

Hu, Y. and Schennach, S. M. (2008). Instrumental variable treatment of nonclassical measurement error models. *Econometrica*, 76(1):195–216.

Huber, P. (2004). *Robust Statistics*. Wiley Series in Probability and Statistics - Applied Probability and Statistics Section Series. Wiley.

Javanmard, A. and Montanari, A. (2018). Debiasing the lasso: Optimal sample size for gaussian designs. *The Annals of Statistics*, 46(6A):2593–2622.

Jordan, M. I., Lee, J. D., and Yang, Y. (2019). Communication-efficient distributed statistical inference. *Journal of the American Statistical Association*, 114(526):668–681.

Kemp, G. C. and Santos-Silva, J. (2012). Regression towards the mode. *Journal of Econometrics*, 170(1):92–101.

Khardani, S. and Yao, A. (2017). Non linear parametric mode regression. *Communications in Statistics-Theory and Methods*, 46(6):3006–3024.

Khuri, A. I., Mukherjee, B., Sinha, B. K., and Ghosh, M. (2006). Design issues for generalized linear models: A review. *Statistical Science*, 21(3):376–399.

Kiefer, J. (1959). Optimum experimental designs. *Journal of the Royal Statistical Society: Series B (Methodological)*, 21(2):272–304.

Knight, K. (1998). Limiting distributions for l1 regression estimators under general conditions. *Annals of Statistics*, pages 755–770.

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

Koenker, R. (2017). Quantile regression: 40 years on. *Annual Review of Economics*, 9:155–176.

Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrica*, 46(1):33–50.

Krief, J. M. (2017). Semi-linear mode regression. *The Econometrics Journal*, 20(2):149–167.

Leadbetter, M., Lindgren, G., and Rootzén, H. (1983). *Extremes and Related Properties of Random Sequences and Processes*. Springer.

Lee, J. D., Liu, Q., Sun, Y., and Taylor, J. E. (2017). Communication-efficient sparse regression. *The Journal of Machine Learning Research*, 18(1):115–144.

Lee, M.-J. (1989). Mode regression. *Journal of Econometrics*, 42(3):337–349.

Lee, M.-J. (1993). Quadratic mode regression. *Journal of Econometrics*, 57(1-3):1–19.

Lee, M.-J. and Kim, H. (1998). Semiparametric econometric estimators for a truncated regression model: A review with an extension. *Statistica Neerlandica*, 52(2):200–225.

Ma, P., Mahoney, M. W., and Yu, B. (2015). A statistical perspective on algorithmic leveraging. *The Journal of Machine Learning Research*, 16(1):861–911.

Manski, C. F. (1991). Regression. *Journal of Economic Literature*, 29(1):34–50.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models, Second Edition*. Chapman and Hall/CRC Monographs on Statistics and Applied Probability Series. Chapman & Hall.

Nesterov, Y. (2018). *Lectures on convex optimization*, volume 137. Springer.

Newey, W. and McFadden, D. (1986). Large sample estimation and hypothesis testing. In Engle, R. F. and McFadden, D., editors, *Handbook of Econometrics*, volume 4, chapter 36, pages 2111–2245. Elsevier, 1 edition.

Ohlsson, E. (1989). Asymptotic normality for two-stage sampling from a finite population. *Probability Theory and Related Fields*, 81(3):341–352.

Ota, H., Kato, K., and Hara, S. (2019). Quantile regression approach to conditional mode estimation. *Electronic Journal of Statistics*, 13(2):3120–3160.

Parzen, M. I., Wei, L. J., and Ying, Z. (1994). A resampling method based on pivotal estimating equations. *Biometrika*, 81:341–350.

Politis, D., Romano, J., and Wolf, M. (1999). *Subsampling*. Springer.

Portnoy, S. and Koenker, R. (1997). The gaussian hare and the laplacian tortoise: computability of squared-error versus absolute-error estimators. *Statistical Science*, 12(4):279–300.

Powell, J. L. (1986). Censored regression quantiles. *Journal of Econometrics*, 32(1):143–155.

Pukelsheim, F. (2006). *Optimal design of experiments*. SIAM.

Raab, M. and Steger, A. (1998). "balls into bins" - a simple and tight analysis. In *Proceedings of the Second International Workshop on Randomization and Approximation Techniques in Computer Science*, RANDOM '98, page 159170, Berlin, Heidelberg. Springer-Verlag.

Raskutti, G. and Mahoney, M. W. (2016). A statistical perspective on randomized sketching for ordinary least-squares. *The Journal of Machine Learning Research*, 17(1):7508–7538.

Reiman, D. M. and Göhre, B. E. (2019). Deblending galaxy superpositions with branched generative adversarial networks. *Monthly Notices of the Royal Astronomical Society*, 485(2):2617–2627.

Rockafellar, R. T. (1970). *Convex Analysis*. Princeton Mathematical Series. Princeton University Press, Princeton, N. J.

Rousseeuw, P. J. and Hubert, M. (2011). Robust statistics for outlier detection. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(1):73–79.

Rudelson, M. (1999). Random vectors in the isotropic position. *Journal of Functional Analysis*, 164:60–72.

Rudelson, M. and Zhou, S. (2012). Reconstruction from anisotropic random measurements. In *Conference on Learning Theory*, pages 10–1.

Ruggles, S., Flood, S., Goeken, R., Grover, J., Meyer, E., Pacas, J., and Sobek, M. (2020). Ipums usa: Version 10.0 [dataset]. minneapolis, mn: Ipums; 2020.

Ruppert, D. and Carroll, R. J. (1980). Trimmed least squares estimation in the linear model. *Journal of the American Statistical Association*, 75(372):828–838.

Sager, T. W. and Thisted, R. A. (1982). Maximum likelihood estimation of isotonic modal regression. *Ann. Statist.*, 10(3):690–707.

Sasaki, H., Ono, Y., and Sugiyama, M. (2016). Modal regression via direct log-

density derivative estimation. In *International Conference on Neural Information Processing*, pages 108–116.

Shamir, O., Srebro, N., and Zhang, T. (2014). Communication-efficient distributed optimization using an approximate newton-type method. In *International conference on machine learning*, pages 1000–1008.

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

Ting, D. and Brochu, E. (2018). Optimal subsampling with influence functions. In *Advances in Neural Information Processing Systems*, pages 3650–3659.

Van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202.

van der Vaart, A. W. (2000). *Asymptotic Statistics*, volume 3. Cambridge University Press.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak convergence and empirical processes: with applications to statistics*. Springer.

Vert, J.-P. (2015). *apg: Optimization with accelerated proximal gradient*. R package version 0.1.1.

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

Wang, H., Yang, M., and Stufken, J. (2019). Information-based optimal subdata

selection for big data linear regression. *Journal of the American Statistical Association*, 114(525):393–405.

Wang, H., Zhu, R., and Ma, P. (2018). Optimal subsampling for large sample logistic regression. *Journal of the American Statistical Association*, 113(522):829–844.

Wang, J., Kolar, M., Srebro, N., and Zhang, T. (2017a). Efficient distributed learning with sparsity. In *International Conference on Machine Learning*, pages 3636–3645.

Wang, X., Chen, H., Cai, W., Shen, D., and Huang, H. (2017b). Regularized modal regression with applications in cognitive impairment prediction. In *Advances in Neural Information Processing Systems 30*, pages 1448–1458.

Wang, Y., Yu, A. W., and Singh, A. (2017c). On computationally tractable selection of experiments in measurement-constrained regression models. *The Journal of Machine Learning Research*, 18(1):5238–5278.

Wasserman, L. (2006). *All of nonparametric statistics*. Springer Science & Business Media.

Wei, Y., Kehm, R. D., Goldberg, M., and Terry, M. B. (2019). Applications for quantile regression in epidemiology. *Current Epidemiology Reports*, 6(2):191–199.

Western, B. and Rosenfeld, J. (2011). Unions, norms, and the rise in us wage inequality. *American Sociological Review*, 76(4):513–537.

Williams, R. (2012). Using the margins command to estimate and interpret adjusted predictions and marginal effects. *The Stata Journal*, 12(2):308–331.

Xu, G., Sit, T., Wang, L., and Huang, C.-Y. (2017). Estimation and inference of quantile regression for survival data under biased sampling. *Journal of the American Statistical Association*, 112(520):1571–1586.

Xu, P., Yang, J., Roosta, F., Ré, C., and Mahoney, M. W. (2016). Sub-sampled newton methods with non-uniform sampling. In *Advances in Neural Information Processing Systems*, pages 3000–3008.

Yao, W. and Li, L. (2014). A new regression model: modal linear regression. *Scandinavian Journal of Statistics*, 41(3):656–671.

Yao, W., Lindsay, B., and Li, R. (2012). Local modal regression. *Journal of Nonparametric Statistics*, 24(3):647–663.

Yao, W. and Xiang, S. (2016). Nonparametric and varying coefficient modal regression. *arXiv preprint arXiv:1602.06609*.

Yu, Y., Chao, S.-K., and Cheng, G. (2021). Distributed bootstrap for simultaneous inference under high dimensionality. *arXiv preprint arXiv:2102.10080*.

Zhan, X. (2004). *Matrix Inequalities*. Springer.

Zhang, A., Brown, L. D., and Cai, T. T. (2016). Semi-supervised inference: General theory and estimation of means. *arXiv preprint arXiv:1606.07268*.

Zhang, C.-H. and Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, pages 217–242.

Zhang, T., Kato, K., and Ruppert, D. (2021+). Bootstrap inference for quantile-based modal regression. *Journal of the American Statistical Association*, forthcoming.

Zheng, Q., Peng, L., and He, X. (2015). Globally adaptive quantile regression with ultra-high dimensional data. *Annals of Statistics*, 43(5):2225.

Zhu, X. J. (2005). Semi-supervised learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences.