UNIVERSITY OF CALIFORNIA SAN DIEGO

Scalable Quantile Learning

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Mathematics with a Specialization in Statistics

by

Xiaoou Pan

Committee in charge:

Professor Wenxin Zhou, Chair
Professor Ery Arias-Castro
Professor Dimitris N. Politis
Professor Jason Schweinsberg
Professor Yixiao Sun

2022

The Dissertation of Xiaoou Pan is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2022

DEDICATION

To my confusion and hesitation along this journey.

EPIGRAPH

There is only one heroism in the world:
to see the world as it is, and to love it.

Romain Rolland

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First and foremost, I would like to express my greatest appreciation to my advisor, collaborator and friend Wenxin Zhou. He brought the topic of this dissertation into my attention and offered invaluable help, academically and non-academically, over the past five years. I still remember in my first project, Wenxin considerately decomposed the burdensome proofs into small pieces, so that it became easier for me to solve them. This work would not have been finished without his guidance.

Besides, I feel very fortunate to have met every member in my committee. Prof. Ery Arias-Castro was my first-year faculty advisor, and his support in my first summer substantially promoted my research progress. Prof. Jason Schweinsberg is my role model at UCSD. His passion for teaching and research has always been inspiring me. Through my advancement talk and other discussions, Prof. Dimitris Politis and Prof. Yixiao Sun provided me valuable insights and advice on bootstrap and econometrics, which considerably improved the quality of this work.

Along my academic journey, I have been honored and humbled to work with several talented scholars. Prof. Zhiliang Ying advised my undergraduate thesis, and led me to the world of mathematical statistics. I owe my gratitude to Prof. Xuming He and Prof. Kean Ming Tan, with whom I completed at least two papers, and received warm care and encouragement. When I was at University of Michigan, Prof. Jian Kang and Prof. Ji Zhu introduced the topics of high-dimensional statistics and statistical computing to me that paved the way for my research. In 2021, I spent a wonderful summer at Mayo Clinic as an intern working with Vivien Yin, and I learnt Bayesian statistics and causal inference under her mentorship.

I also wish to thank my friends and classmates at UCSD, too many to name, for those trips to La Jolla beach and downtown Dan Diego. Finally, my special thanks go to my parents Weihuai and Lili. Their love and unconditional support is the lighthouse guiding me through thick and thin.

Chapter 2, in part, is a reprint of the material in the paper "Multiplier bootstrap for quantile regression: Non-asymptotic theory under random design", Pan, Xiaoou and Zhou, Wen-Xin. The paper has been published on *Information and Inference: A Journal of the IMA*, **10** 813-861. The dissertation author was the primary investigator and author of this paper.

Chapter 3, in part, is a reprint of the material in the paper "Smoothed quantile regression with large-scale inference", He, Xuming; Pan, Xiaoou; Tan, Kean Ming and Zhou, Wen-Xin. The paper has been accepted by *Journal of Econometrics*, 2022. The dissertation author was the primary investigator and author of this paper.

Chapter 4, in part, is a reprint of the material in the paper "Scalable estimation and inference for censored quantile regression process", He, Xuming; Pan, Xiaoou; Tan, Kean Ming and Zhou, Wen-Xin. The paper has been reviewed by *Annals of Statistics* and is currently under major revision. The dissertation author was the primary investigator and author of this paper.

VITA

2014	B.S. in Mathematics, Fudan University
2016	M.S. in Statistics, University of Michigan Ann Arbor
2017–2022	Graduate Teaching and Research Assistant, University of California San Diego
2022	Ph. D. in Mathematics with a Specialization in Statistics, University of California San Diego

PUBLICATIONS

MAN, R., PAN, X., TAN, K, M, and ZHOU, W.-X., A Unified Algorithm for Penalized Convolution Smoothed Quantile Regression (2022), *Journal of Computational and Graphical Statistics*, submitted

HE, X., PAN, X., TAN, K, M, and ZHOU, W.-X., Scalable estimation and inference for censored quantile regression process (2022), *Annals of Statistics*, under revision

HE, X., PAN, X., TAN, K, M, and ZHOU, W.-X., Smoothed quantile regression with large-scale inference (2022), *Journal of Econometrics*, accepted

PAN, X. and ZHOU, W.-X., Multiplier bootstrap for quantile regression: Non-asymptotic theory under random design (2021), *Information and Inference: A Journal of the IMA*, **10** 813-861

PAN, X., SUN, Q. and ZHOU, W.-X., Iteratively reweighted ℓ_1-penalized robust regression (2021), *Electronic Journal of Statistics*, **15** 3287-3348

BOSE, K., FAN, J., KE, Y., PAN, X. and ZHOU, W.-X., FarmTest: An R package for factor-adjusted robust multiple testing (2020), *The R Journal*, **12** 372–387

ABSTRACT OF THE DISSERTATION

Scalable Quantile Learning

by

Xiaoou Pan

Doctor of Philosophy in Mathematics with a Specialization in Statistics

University of California San Diego, 2022

Professor Wenxin Zhou, Chair

Quantile regression (QR) is a powerful tool for learning the relationship between a continuous outcome and a set of covariates while exploring heterogeneous effects. This dissertation focuses on statistical learning (estimation and inference) in the increasing dimensional regime with random designs, and the outcome is possibly subject to random censoring. We provide a comprehensive analysis on three problems: (i) the classical QR and multiplier bootstrap inference; (ii) QR with a convolution-based smoothed approach that achieves adequate approximation to computation and inference; (iii) censored QR with a smoothed martingale-based sequential estimating equations approach, and an ℓ_1-regularized regression problem in the high-dimensional regime. The unified principle of these methods is to turn the non-differentiable check function

into a twice-differentiable, globally convex and locally strongly convex surrogate, which admits fast and scalable gradient-based algorithms to perform optimization. For all the aforementioned tasks, we theoretically establish explicit non-asymptotic bounds on estimation and Bahadur-Kiefer linearization errors, from which we show that the asymptotic normality holds, when the covariate dimension grows with the sample size at a sublinear rate. In particular, uniform convergence rate (over a range of quantile indexes) and weak convergence are established for censored quantile regression process. The multiplier bootstrap inference, as a companion, is also rigorously justified for all the problems. Extensive numerical experiments confirm the computational scalability and reliability to large-scale data, and demonstrate the advantage of our methods over existing ones.

Chapter 1

Introduction

1.1 Quantile Regression

Since Koenker and Bassett's seminal work [Koenker and Bassett, 1978], quantile regression (QR) has attracted enormous attention in statistics, econometrics and other scientific fields. Compared to the least squares regression that focuses on modeling the conditional mean of y given $\boldsymbol{x}$, quantile regression models the entire conditional distribution of y given $\boldsymbol{x}$, and thus provides valuable insights into heterogeneity in the relationship between $\boldsymbol{x}$ and y. Moreover, quantile regression is robust against outliers and can be performed for skewed or heavy-tailed response distributions without correct specification of the likelihood. These advantages make quantile regression an appealing method to explore data features that are invisible to the least squares regression.

Classical theory of quantile regression including statistical consistency and asymptotic normality has been thoroughly developed [Bassett and Koenker, 1978, 1986, Portnoy and Koenker, 1989, Welsh, 1989, Pollard, 1991, Zhao, Rao and Chen, 1993, Arcones, 1996, He and Shao, 1996, 2000]. A common thread of the previous work is that the regression estimators are studied under the fixed design setting, that is, the covariates $\{\boldsymbol{x}_i\}_{i=1}^n$ are deterministic vectors with a fixed dimension and satisfy some (asymptotic and non-asymptotic) conditions, and the only randomness arises from the regression errors $\{\varepsilon_i\}_{i=1}^n$. A comprehensive review of the asymptotic theory under fixed design can be found in Chapter 4 of Koenker [2005].

In contrast to fixed designs, modern statistics have emphasized non-asymptotic results in the random design setting, where the covariates $\{\boldsymbol{x}_i\}_{i=1}^n$ are treated as random vectors [Vershynin, 2018, Wainwright, 2019], and the dimension $p = p_n$ is subject to a growth condition $p \asymp n^a$ for some $a \in (0,1)$. This additional randomness increases the complexity of the model, and makes theoretical analysis more subtle because the empirical processes involved now depend on the random covariates with dimensionality possibly growing with the sample size. A main difficulty is that the quantile loss is piecewise linear, and hence its "curvature energy" is concentrated in a single point. This is substantially different from other popular regression loss functions, such as the least squared loss and Huber loss, which are at least locally strongly convex. The lack of smoothness and strong convexity makes it much more challenging to establish non-asymptotic theory for quantile regression under random designs, and is the main motivation of this dissertation. We refer to Belloni and Chernozhukov [2011], Chao, Volgushev and Cheng [2017], Belloni et al. [2019] for some well-known developments of quantile regression under random designs with growing dimensionality.

1.2 Statistical Inference for Quantile Regression

In addition to the finite sample theory of standard quantile regression, we are also interested in large-scale inference for quantile regression under the increasing dimension regime. Broadly speaking, inference of quantile regression can be categorized into two classes: normal calibration and bootstrap calibration (resampling) methods. Normal calibration heavily depends on either the estimation of $1/f_{\varepsilon|\boldsymbol{x}}(0)$, where $f_{\varepsilon|\boldsymbol{x}}(\cdot)$ is the conditional density function of ε given $\boldsymbol{x}$, or the regression rank scores [Gutenbrunner and Jurečková, 1992]. Even if the asymptotic variance is well estimated, its approximation accuracy to the finite-sample variance depends on the design matrix and the quantile level. Resampling, or bootstrap calibration methods [Efron, 1979], on the other hand, provide a more reliable approach to quantile regression inference. Over the past two decades, various bootstrap calibration methods have been developed for constructing

confidence intervals, including the residual bootstrap and pairwise bootstrap (Section 9.5 of Efron and Tibshirani [1994]), bootstrapping pivotal estimation functions method [Parzen, Wei and Ying, 1994], Markov chain marginal bootstrap [He and Hu, 2002, Kocherginsky, He and Mu, 2005] and wild bootstrap [Feng, He and Hu, 2011]. Inevitably, the resampling approach requires repeatedly computing QR estimates up to thousands of times, and therefore is unduly expensive for large-scale data. For relatively small samples or in the presence of heteroscedastic errors, resampling methods have proven to outperform calibration through the normal approximation. Therefore, in this dissertation we only focus on the resampling method.

Among a variety of bootstrap methods, we are primarily interested in the multiplier bootstrap, also known as the weighted bootstrap, which is one of the most widely used inference tools for constructing confidence intervals and measuring the significance of a test. The theoretical validity of the empirical bootstrap [Efron, 1979] is typically guaranteed by the bootstrapped law of large numbers and central limit theorem; see, for example, Giné and Zinn [1990], Arcones and Giné [1992], Praestgaard and Wellner [1993] and Wellner and Zhan [1996], among others. Rigorous theoretical guarantees of the multiplier bootstrap for M-estimation can be found in Chatterjee and Bose [2005] and Ma and Kosorok [2005], in which $\sqrt{n}$-consistency and asymptotic normality are established. See also Cheng and Huang [2010] for extensions to general semi-parametric models. It has since become an effective and nearly universal inference tool for both parametric and semi-parametric M-estimations. We refer to Spokoiny and Zhilova [2015] for the use of multiplier bootstrap on constructing likelihood-based confidence sets, and Chen and Zhou [2020] for a systematic study of multiplier bootstrap for adaptive Huber regression [Sun, Zhou and Fan, 2020] with applications to large-scale multiple testing for heavy-tailed data.

1.3 Computation of Quantile Regression

Quantile regression involves a convex optimization problem with a piecewise linear loss function, also known as the check function. One can reformulate the QR problem as a linear

program (LP), solvable by the Frisch-Newton algorithm with an average-case computational complexity that grows as a cubic function of p, i.e., $\mathcal{O}_{\mathbb{P}}(n^{1+\alpha}p^3 \log n)$ for some constant $\alpha \in (0, 1/2)$ [Portnoy and Koenker, 1997], where n is the sample size and p is the parametric dimension. However, when applied to large-scale problems—both n and p are large, QR computation via LP reformulation tends to be slow or too memory-intensive. To better appreciate such a challenge, we take the empirical study of U.S. equities from Gu, Kelly and Xiu [2020] as an example. The dataset consists of monthly total individual equity returns, which begins in March 1957 and ends in December 2016, from CRSP for all firms listed in the NYSE, AMEX, and NASDAQ. Within this span of 60 years, the average number of stocks considered is around 6200 per month. After processing, the number of observations (over 60 years) in the entire panel exceeds 4 million, and the number of stock-level covariates is 920. Even with preprocessing, the interior point QR solver in `R` [Koenker, 2022] may either run out of memory or take too much time on a personal computer. This shortcoming arguably makes QR less attractive compared to various machine learning tools. Chapter 5 of Koenker et al. [2017] provides an overview of the prevailing computational methods for quantile regression, such as simplex-based algorithms [Barrodale and Roberts, 1974, Koenker and d'Orey, 1987], interior point methods [Portnoy and Koenker, 1997], and alternating direction method of multipliers, among other first-order proximal methods [Parikh and Boyd, 2014].

1.4 Censored Quantile Regression

Censored data are prevalent in many scientific areas where the response variable of interest is partially observed, mostly due to loss of follow-up. For instance, in a lung cancer study considered by Shedden et al. [2008], 46.6% of the lung cancer patients' survival time are censored, due to either early withdrawal from the study or death because of other reasons that are unrelated to lung cancer. Commonly used methods to study the association between the censored response and explanatory variables (covariates) are through the use of Cox proportional

hazards model and the accelerated failure time model [Andersen et al., 1993, Kleinbaum and Klein, 2012]. Both models assume homogeneous covariate effects and are not applicable to the case in which the lower and upper quantiles of the conditional distribution of the censored response, potentially with different covariate effects, are of interest. Moreover, in many scientific studies, the quantiles of the censored response are of more interest than the mean effect. To capture heterogeneous covariate effects and to better predict the censored response at different quantile levels, various censored quantile regression (CQR) methods have been developed under different assumptions on the censoring mechanism [Powell, 1984, 1986, Ying, Jung and Wei, 1995, Buchinsky and Hahn, 1998, Chernozhukov and Hong, 2002, Honoré, Khan and Powell., 2002, Portnoy, 2003, Wang and Wang, 2009, Leng and Tong, 2013, Yang, Narisetty and He, 2018, De Backer, El Ghouch and Van Keilegom, 2019, 2020]. In addition, a comprehensive review of censored quantile regression can be found in Chapters 6 and 7 in Koenker et al. [2017] as well as Peng [2021].

In this dissertation, we consider the random right censoring mechanism, in which the censoring points are unknown for the uncensored observations. Statistical methods for CQR were first proposed under the stringent assumption that the uncensored response variable (not observable due to censoring) is marginally independent of the censoring variable; see, for example Ying, Jung and Wei [1995], Honoré, Khan and Powell. [2002]. Under a more relaxed conditional independence assumption, conditioned on the covariates, Portnoy [2003] generalized the Kaplan-Meier estimator for estimating the (univariate) survival function to the regression setting, based on Efron [1967]'s redistribution-of-mass construction. From a different perspective, Peng and Huang [2008] employed a martingale-based approach for fitting CQR, and the resulting method has been shown to be closely related to Portnoy [2003]'s method [Neocleous, Branden and Portnoy, 2006, Peng, 2012]. Both Portnoy [2003]'s and Peng and Huang [2008]'s methods, along with their variants, involve solving a series of quantile regression problems that can be reformulated as linear programs, solvable by the simplex or interior point method [Barrodale and Roberts, 1974, Portnoy and Koenker, 1997, Koenker and Mizera, 2014]. Statistical properties of

the aforementioned methods have been well studied, assuming that the number of covariates, p, is fixed [Neocleous, Branden and Portnoy, 2006, Peng and Huang, 2008, Portnoy and Lin, 2010, Peng, 2012]. To this date, the impact of dimensionality in the increasing-p regime, in which p is allowed to increase with the number of observations, remains unclear in the presence of censored outcomes.

In the high-dimensional setting in which $p > n$, convex and nonconvex penalty functions are often employed to perform variable selection to achieve a trade-off between statistical bias and model complexity. While penalized Cox proportional hazards and accelerated failure time models have been well studied [Fan and Li, 2002, Huang, Ma and Xie, 2006, Cai, Huang and Tian, 2009, Bradic, Fan and Jiang, 2011], existing work on penalized CQR under the framework of Portnoy [2003] and Peng and Huang [2008] in the high-dimensional setting is relatively lacking. Large-sample properties of penalized CQR estimators were first derived under the fixed-p setting ($p < n$), mainly due to the technical challenges introduced by the recursive nature of the procedure [Shows, Lu and Zhang, 2010, Wang, Zhou and Li, 2013, Volgushev, Vagener and Dette, 2014]. More recently, Zheng, Peng and He [2018] studied a penalized CQR estimator, extending the method of Peng and Huang [2008] to the high-dimensional setting $(p > n)$. They showed that the estimation error (under ℓ_2-norm) of the ℓ_1-penalized CQR estimator is upper bounded by $\mathscr{O}\big(\exp(Cs)\sqrt{s\log(p)/n}\big)$, where $C > 0$ is a dimension-free constant. Compared to ℓ_1-penalized QR for uncensored data [Belloni and Chernozhukov, 2011], whose convergence rate is of order $\mathscr{O}\big(\sqrt{s\log(p)/n}\big)$, there is a substantial gap in terms of the impact of the sparsity parameter s.

In addition to the above theoretical issues, our study is also motivated by the computational hardness of CQR under the framework of Portnoy [2003] and Peng and Huang [2008] for problems with large dimension. Their framework involves fitting a series of quantile regressions sequentially over a dense grid of quantile indexes, each of which is solvable by the Frisch-Newton algorithm with computational complexity that grows as a cubic function of p [Portnoy and Koenker, 1997]. Moreover, under the regime in which $p < n$, the asymptotic

covariance matrix of the estimator is rather complicated and thus resampling methods are often used to perform statistical inference [Portnoy, 2003, Peng and Huang, 2008]. A sample-based inference procedure (without resampling) for Peng-Huang's estimator [Peng and Huang, 2008] is available by adapting the plug-in covariance estimation method from Sun et al. [2016]. In the high-dimensional setting in which $p > n$, the computation of ℓ_1-penalized QR is based on either reformulation as linear programs [Koenker and Ng, 2005] or alternating direction method of multiplier algorithms [Yu, Lin and Wang, 2017, Gu et al., 2018]. These algorithms are generic and applicable to a broad spectrum of problems but lack scalability. Since ℓ_1-penalized CQR not only requires the estimation of the whole quantile regression process, but also relies on cross-validation to select the sequence of (mostly different) penalty levels, the state-of-the-art methods [Zheng, Peng and He, 2018, Fei et al., 2021] can be highly inefficient when applied to large-p problems.

1.5 Organization

The rest of this dissertation is organized as follows. In Chapter 2, we formulate the quantile regression problem, then develop estimation theory and justify multiplier bootstrap inference. In Chapter 3, we provide a comprehensive study on quantile regression with a convolution-type smoothing mechanism, and develop a scalable computational device. In Chapter 4, we analyze (regularized) censored quantile regression via a smoothed estimation equation approach, and establish theory as well as computational methods. Throughout the dissertation, we focus on non-asymptotic theory with growing (intrinsic) dimension under random designs. All the theoretical proofs are collected in the Appendix.

1.6 Notation

For every integer $k \geq 1$, we use $\mathbb{R}^k$ to denote the the k-dimensional Euclidean space. The inner product of any two vectors $\boldsymbol{u} = (u_1, \ldots, u_k)^{\mathrm{T}}, \boldsymbol{v} = (v_1, \ldots, v_k)^{\mathrm{T}} \in \mathbb{R}^k$ is defined by $\boldsymbol{u}^{\mathrm{T}}\boldsymbol{v} =$

$\langle \boldsymbol{u}, \boldsymbol{v} \rangle = \sum_{i=1}^{k} u_i v_i$. We use $\|\cdot\|_p$ $(1 \le p \le \infty)$ to denote the ℓ_p-norm in $\mathbb{R}^k$: $\|\boldsymbol{u}\|_p = (\sum_{i=1}^{k} |u_i|^p)^{1/p}$ and $\|\boldsymbol{u}\|_\infty = \max_{1 \le i \le k} |u_i|$. For $k \ge 2$, $\mathbb{S}^{k-1} = \{\boldsymbol{u} \in \mathbb{R}^k : \|\boldsymbol{u}\|_2 = 1\}$ denotes the unit sphere in $\mathbb{R}^k$. Throughout the dissertation, we use bold capital letters to represent matrices. For $k \ge 2$, $\mathbf{I}_k$ represents the identity/unit matrix of size k. For any $k \times k$ symmetric matrix $\mathbf{A} \in \mathbb{R}^{k \times k}$, $\|\mathbf{A}\|_2$ is the operator norm of $\mathbf{A}$, and we use $\underline{\lambda}_{\mathbf{A}}$ and $\overline{\lambda}_{\mathbf{A}}$ to denote the minimal and maximal eigenvalues of $\mathbf{A}$, respectively. For a positive semidefinite matrix $\mathbf{A} \in \mathbb{R}^{k \times k}$, $\|\cdot\|_{\mathbf{A}}$ denotes the norm linked to $\mathbf{A}$ given by $\|\boldsymbol{u}\|_{\mathbf{A}} = \|\mathbf{A}^{1/2}\boldsymbol{u}\|_2$, $\boldsymbol{u} \in \mathbb{R}^k$. Moreover, given $r \ge 0$, define the Euclidean ball and ellipse as $\mathbb{B}^k(r) = \{\boldsymbol{u} \in \mathbb{R}^k : \|\boldsymbol{u}\|_2 \le r\}$ and $\mathbb{B}_{\mathbf{A}}(r) = \{\boldsymbol{u} \in \mathbb{R}^k : \|\boldsymbol{u}\|_{\mathbf{A}} \le r\}$, respectively. For any integer $d \ge 1$, we write $[d] = \{1, \ldots, d\}$. For any set $\mathscr{S}$, we use $|\mathscr{S}|$ to denote its cardinality, i.e. the number of elements in $\mathscr{S}$. Given an event/subset $\mathscr{A}$, $\mathbb{1}\{\mathscr{A}\}$ or $\mathbb{1}_{\mathscr{A}}$ represents the indicator function of this event/subset. For any two real numbers a and b, we write $a \wedge b = \min\{a, b\}$ and $a \vee b = \max\{a, b\}$. For two sequences of non-negative numbers $\{a_n\}_{n \ge 1}$ and $\{b_n\}_{n \ge 1}$, $a_n \lesssim b_n$ indicates that there exists a constant $C > 0$ independent of n such that $a_n \le C b_n$; $a_n \gtrsim b_n$ is equivalent to $b_n \lesssim a_n$; $a_n \asymp b_n$ is equivalent to $a_n \lesssim b_n$ and $b_n \lesssim a_n$.

Chapter 2

Quantile Regression: A Finite Sample Perspective

2.1 Theory for Estimation and Inference

2.1.1 Finite sample theory under random design

We consider a response variable y and p-dimensional covariates $\boldsymbol{x} = (x_1, \ldots, x_p)^\intercal$ such that the τ-th ($0 < \tau < 1$) conditional quantile of y given $\boldsymbol{x}$ is given by $F^{-1}_{y|\boldsymbol{x}}(\tau|\boldsymbol{x}) = \langle \boldsymbol{x}, \boldsymbol{\beta}^* \rangle$, where $\boldsymbol{\beta}^* = (\beta^*_1, \ldots, \beta^*_p)^\intercal \in \mathbb{R}^p$. Here we assume $x_1 \equiv 1$ so that β^*_1 represents the intercept. Let $\{(y_i, \boldsymbol{x}_i)\}_{i=1}^n$ be independent and identically distributed (iid) data vectors from $(y, \boldsymbol{x})$. The preceding model assumption is equivalent to

$$y_i = \langle \boldsymbol{x}_i, \boldsymbol{\beta}^* \rangle + \varepsilon_i, \tag{2.1}$$

where ε_i's are independent noise variables that satisfy $\mathbb{P}(\varepsilon_i \leq 0 \,|\, \boldsymbol{x}_i) = \tau$. The quantile regression estimator of $\boldsymbol{\beta}^*$ is then defined as

$$\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}(\tau) \in \operatorname*{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^p} Q_n(\boldsymbol{\beta}), \tag{2.2}$$

where

$$Q_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \rho_\tau(y_i - \langle \boldsymbol{x}_i, \boldsymbol{\beta}\rangle) \quad \text{with} \quad \rho_\tau(u) = u\{\tau - I(u<0)\} \tag{2.3}$$

is the empirical loss. The loss function ρ_τ is known as the "check function" or "pinball loss".

This section presents two non-asymptotic results, the concentration inequality and Bahadur representation, for the quantile regression estimator under random design. First, we specify the conditions on the random pair $(\boldsymbol{x}, \varepsilon)$ under which the analysis applies.

Condition 2.1.1 (Random design)**.** *The random predictor $\boldsymbol{x} \in \mathbb{R}^p$ is sub-Gaussian: there exists $\upsilon_0 \geq 1$ such that $\mathbb{P}(|\langle \boldsymbol{u}, \boldsymbol{x}\rangle| \geq \upsilon_0 \|\boldsymbol{u}\|_{\boldsymbol{\Sigma}} \cdot t) \leq 2e^{-t^2/2}$ for all $\boldsymbol{u} \in \mathbb{R}^p$ and $t \geq 0$, where $\boldsymbol{\Sigma} = \mathbb{E}(\boldsymbol{x}\boldsymbol{x}^\intercal)$.*

Condition 2.1.2 (Regularity condition on error distribution)**.** *The conditional probability density function of ε given $\boldsymbol{x}$, $f_{\varepsilon|\boldsymbol{x}}(\cdot)$, is continuous on its support. Moreover, there exist constants $\bar{f} \geq \underline{f} > 0$ and $L_0 > 0$ such that*

$$\underline{f} \leq f_{\varepsilon|\boldsymbol{x}}(0) \leq \bar{f} \quad \textit{and} \quad |f_{\varepsilon|\boldsymbol{x}}(u) - f_{\varepsilon|\boldsymbol{x}}(0)| \leq L_0|u| \ \textit{ for all } u \in \mathbb{R}, \ \textit{ almost surely.}$$

Condition 2.1.1 is satisfied for a class of multivariate distributions. Typical examples include: (i) Multivariate Gaussian and (symmetric) Bernoulli distributions, (ii) uniform distribution on the sphere in $\mathbb{R}^p$ with center at the origin and radius $\sqrt{p}$, (iii) uniform distribution on the Euclidean ball, and (iv) uniform distribution on the unit cube $[-1,1]^p$. The constant υ_0 is dimension-free, and thus can be viewed as an absolute constant. See Chapter 6 in Wainwright [2019] and references therein for further discussion of sub-Gaussian distributions in higher dimensions. Condition 2.1.2 on the conditional density function of ε given $\boldsymbol{x}$ is standard and routinely used in the study of quantile regression.

Our first two results characterize the non-asymptotic versions of (i) deviation bound, and (ii) the Bahadur representation for the quantile regression estimator.

Theorem 2.1.1. *Assume Conditions 2.1.1 and 2.1.2 hold. Then, for any $t \geq 0$, the quantile regression estimator $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}(\tau)$ $(0 < \tau < 1)$ given in* (2.2) *satisfies*

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma}} \lesssim \frac{1}{\underline{f}} \sqrt{\frac{p+t}{n}} \tag{2.4}$$

with probability at least $1-2e^{-t}$ as long as $n \gtrsim L_0^2 \underline{f}^{-4}(p+t)$.

Theorem 2.1.2. *Under the same conditions in Theorem 2.1.1, for any $t \geq 0$,*

$$\begin{aligned} &\left\| \mathbf{S}^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) + \mathbf{S}^{-1/2} \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i \{ I(\varepsilon_i \leq 0) - \tau \} \right\|_2 \\ &\lesssim \frac{(p+t)^{1/4}(p\log n + t)^{1/2}}{n^{3/4}} + \frac{(p+\log n)^{1/2} p \log n + (p\log n)^{1/2} t}{n} \end{aligned} \tag{2.5}$$

with probability at least $1-4e^{-t}$ whenever $n \gtrsim L_0^2 \underline{f}^{-4}(p+t)$, where $\mathbf{S} = \mathbb{E}\{f_{\varepsilon|\boldsymbol{x}}(0)\boldsymbol{x}\boldsymbol{x}^\intercal\}$.

The significance of Bahadur representation lies in expression of a complicated nonlinear estimator as a normalized sum of independent random variables from which asymptotically normal behavior follows. To validate this point, the following result provides a Berry-Esseen bound for any linear contrast of the quantile regression estimator.

Theorem 2.1.3. *Let $\boldsymbol{\lambda} \in \mathbb{R}^p$ be a deterministic vector that defines a linear contrast of interest. Under the conditions of Theorem 2.1.1, it holds that*

$$\sup_{x \in \mathbb{R}} \left| \mathbb{P}\big(n^{1/2} \langle \boldsymbol{\lambda}, \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \rangle \leq x \big) - \Phi(x/\sigma_\tau) \right| \lesssim \frac{(p+\log n)^{1/4}(p\log n)^{1/2}}{n^{1/4}}, \tag{2.6}$$

where $\sigma_\tau^2 = \tau(1-\tau)\|\mathbf{S}^{-1}\boldsymbol{\lambda}\|_{\boldsymbol{\Sigma}}^2$ and $\Phi(\cdot)$ denotes the standard normal distribution function.

Remark 2.1.1 (Large-p asymptotics)**.** *Modern statistical research allow $p = p_n \to \infty$ as $n \to \infty$. Results of quantile regression with increasing p are available in Welsh [1989], He and Shao [2000] and Belloni et al. [2019] when $p = o(n)$, and in Belloni and Chernozhukov [2011], Wang, Wu and Li [2012] and Koenker et al. [2017] for regularized quantile regression when $p \gg n$.*

In particular, Welsh [1989] shows that $p^3(\log n)^2/n \to 0$ *suffices for a normal approximation. This growth condition remains the best known one although under weaker assumptions on the (fixed) design [He and Shao, 2000, Belloni et al., 2019]. To our knowledge, the weakest fixed design assumption is* $\max_{1\le i\le n}\|\boldsymbol{x}_i\|_2^2 = \mathscr{O}(p)$*. In the (sub-Gaussian) random design setting, the obtained non-asymptotic Bahadur representation* (2.5) *with* $t = \log n$ *reads:*

$$n^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) = \mathbf{S}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{\tau - I(\varepsilon_i \le 0)\}\boldsymbol{x}_i + O_{\mathbb{P}}\left\{\frac{p^{3/4}(\log n)^{1/2} + p^{1/2}(\log n)^{3/4}}{n^{1/4}} + \frac{p^{3/2}\log n + p(\log n)^{3/2}}{n^{1/2}}\right\}.$$

Combined with a multivariate central limit theorem [Portnoy, 1986] or Theorem 2.1.3, this shows that the normal approximation holds as long as $p^3(\log n)^2/n \to 0$*, which matches the scaling under fixed design although the proofs are entirely different.*

2.1.2 Multiplier bootstrap and confidence estimation

The multiplier bootstrap method, which dates back to Dudewicz [1992] and Barbe and Bertail [1995], is based on reweighting the summands of empirical loss with random weights. To be specific, let $\mathscr{R}_n = \{e_1, \ldots, e_n\}$ be a sequence of independent Rademacher random variables that are independent of the observed data $\mathscr{D}_n = \{(y_i, \boldsymbol{x}_i)\}_{i=1}^n$. That is, $e_i \in \{-1, 1\}$ and satisfies $\mathbb{P}(e_i = 1) = \mathbb{P}(e_i = -1) = 1/2$. Randomly perturb the empirical loss $Q_n(\boldsymbol{\beta}) = (1/n)\sum_{i=1}^n \rho_\tau(y_i - \langle \boldsymbol{x}_i, \boldsymbol{\beta}\rangle)$ by multiplying its summands with $w_i := e_i + 1$, we obtain the bootstrapped loss function

$$Q_n^\flat(\boldsymbol{\beta}) := \frac{1}{n}\sum_{i=1}^{n} w_i \rho_\tau(y_i - \langle \boldsymbol{x}_i, \boldsymbol{\beta}\rangle), \quad \boldsymbol{\beta} \in \mathbb{R}^p. \tag{2.7}$$

Note that $w_i \in \{0, 2\}$ satisfies $\mathbb{E}(w_i) = 1$ and $\mathrm{var}(w_i) = 1$. Moreover, the bootstrapped loss $Q_n^\flat : \mathbb{R}^p \mapsto [0, \infty)$ is also convex.

Let $\mathbb{E}^*(\cdot) = \mathbb{E}(\cdot\,|\,\mathscr{D}_n)$ and $\mathbb{P}^*(\cdot) = \mathbb{P}(\cdot\,|\,\mathscr{D}_n)$ be the conditional expectation and probability

given $\mathscr{D}_n$, respectively. Then we have $\mathbb{E}^*\{Q_n^\flat(\boldsymbol{\beta})\} = Q_n(\boldsymbol{\beta})$ for any $\boldsymbol{\beta} \in \mathbb{R}^p$. This indicates that the quantile estimator $\widehat{\boldsymbol{\beta}}(\tau) = (\widehat{\beta}_1, \ldots, \widehat{\beta}_p)^{\mathsf{T}}$ in the $\mathscr{D}_n$-world is the target parameter in the bootstrap world:

$$\operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} \mathbb{E}^*\{Q_n^\flat(\boldsymbol{\beta})\} = \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} Q_n(\boldsymbol{\beta}) = \widehat{\boldsymbol{\beta}}(\tau).$$

This simple observation motivates the following multiplier bootstrap estimator:

$$\widehat{\boldsymbol{\beta}}^\flat := \widehat{\boldsymbol{\beta}}^\flat(\tau) \in \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} Q_n^\flat(\boldsymbol{\beta}). \tag{2.8}$$

We refer to Chatterjee and Bose [2005] for a general asymptotic theory for weighted bootstrap for estimating equations, where a class of bootstrap weights is considered. Extensions to semiparametric M-estimation can be found in Ma and Kosorok [2005] and Cheng and Huang [2010].

Let $1-\alpha \in (0,1)$ be a prespecified confidence level. Based on the bootstrap statistic $\widehat{\boldsymbol{\beta}}^\flat = (\widehat{\beta}_1^\flat, , \ldots, \widehat{\beta}_p^\flat)^{\mathsf{T}}$, we consider three methods to construct bootstrap confidence intervals.

(i) (Efron's percentile method). For every $1 \le j \le d$ and $q \in (0,1)$, let $\widehat{\zeta}_{j,q}$ be the (conditional) upper q-quantile of $\widehat{\beta}_j^\flat$, that is,

$$\widehat{\zeta}_{j,q} = \inf\big\{z \in \mathbb{R} : \mathbb{P}^*(\widehat{\beta}_j^\flat > z) \le q\big\}. \tag{2.9}$$

Efron's percentile interval is of the form

$$\mathscr{I}_j^{\mathrm{per}} = [\widehat{\zeta}_{j,1-\alpha/2}, \widehat{\zeta}_{j,\alpha/2}], \quad j = 1, \ldots, p. \tag{2.10}$$

(ii) (Normal interval). The second method is the normal interval:

$$\mathscr{I}_j^{\text{norm}} = [\widehat{\beta}_j - z_{\alpha/2}\widehat{\text{se}}_j^{\text{boot}}, \widehat{\beta}_j + z_{\alpha/2}\widehat{\text{se}}_j^{\text{boot}}], \quad j = 1, \ldots, p, \tag{2.11}$$

where $\widehat{\text{se}}_j^{\text{boot}}$ is the conditional standard deviation of $\widehat{\beta}_j^\flat$ given $\mathscr{D}_n$, and $z_{\alpha/2}$ is the upper $\alpha/2$-quantile of the standard normal distribution.

(iii) (Pivotal interval). The third method, which uses the conditional distribution of $\widehat{\boldsymbol{\beta}}^\flat(\tau) - \widehat{\boldsymbol{\beta}}(\tau)$ to approximate the distribution of the pivot $\widehat{\boldsymbol{\beta}}(\tau) - \boldsymbol{\beta}^*$, is the pivotal interval. Specifically, the $1-\alpha$ bootstrap pivotal confidence intervals for β_j^*'s are

$$\mathscr{I}_j^{\text{piv}} = [2\widehat{\beta}_j - \widehat{\zeta}_{j,\alpha/2}, 2\widehat{\beta}_j - \widehat{\zeta}_{j,1-\alpha/2}], \quad j = 1, \ldots, p. \tag{2.12}$$

In fact, there is a simple connection between the bootstrap pivotal interval and the percentile interval: the percentile interval is the pivotal interval reflected about the point $\widehat{\beta}_j$.

Before we formally investigate the theoretical properties of the bootstrap estimator $\widehat{\boldsymbol{\beta}}^\flat(\tau)$, recall the Bahadur representation of $\widehat{\boldsymbol{\beta}}(\tau)$:

$$\widehat{\boldsymbol{\beta}}(\tau) = \boldsymbol{\beta}^* + \frac{1}{n}\sum_{i=1}^{n}\{\tau - I(\varepsilon_i \le 0)\}\mathbf{S}^{-1}\boldsymbol{x}_i + \boldsymbol{r}_n,$$

where $\boldsymbol{r}_n$ is the higher-order remainder term. Heuristically, the bootstrap estimator $\widehat{\boldsymbol{\beta}}^\flat(\tau)$ can be viewed as the quantile regression estimator of $\widehat{\boldsymbol{\beta}}(\tau)$ in the bootstrap world under the model $y_i = \langle \boldsymbol{x}_i, \widehat{\boldsymbol{\beta}}(\tau)\rangle + \varepsilon_i^\flat$. According to the Bahadur representation, it can be written as $y_i \approx \langle \boldsymbol{x}_i, \boldsymbol{\beta}^*\rangle + (1/n)\sum_{i=1}^n \langle \boldsymbol{x}_i, \mathbf{S}^{-1}\boldsymbol{x}_i\rangle\{\tau - I(\varepsilon_i \le 0)\}$. The accuracy of the percentile interval, however, relies on the property that $\widehat{\boldsymbol{\beta}}_\tau^\flat$ is randomly concentrated around $\boldsymbol{\beta}^*$. Motivated by this observation and the finite-sample correction method used in Feng, He and Hu [2011], for practical implementation we replace the original response y_i in the multiplier bootstrap by $\widehat{y}_i = y_i - \{\widehat{f}_\varepsilon(0)\}^{-1}h_i\{\tau - I(\widehat{\varepsilon}_i \le 0)\}$, where $h_i = \boldsymbol{x}_i^{\mathrm{T}}(\sum_{j=1}^n \boldsymbol{x}_j\boldsymbol{x}_j^{\mathrm{T}})^{-1}\boldsymbol{x}_i$ and $\widehat{f}_\varepsilon(0)$ is estimated from

the fitted residuals $\widehat{\varepsilon}_i = y_i - \langle \boldsymbol{x}_i, \widehat{\boldsymbol{\beta}}(\tau)\rangle$. In particular, the density estimate $\widehat{f_\varepsilon}$ employs the adaptive kernel method [Silverman, 1986].

Back to $\widehat{\boldsymbol{\beta}}^\flat$ defined in (2.8), the following two results provide (i) a conditional deviation inequality, and (ii) a conditional Bahadur representation conditioned on some event that occurs with high probability.

Theorem 2.1.4. *Assume Conditions 2.1.1 and 2.1.2 hold. For any $t \geq 0$, there exists some event $\mathscr{E}_t$ with $\mathbb{P}\{\mathscr{E}(t)\} \geq 1 - 2e^{-t}$ such that the bound* (2.4) *holds on $\mathscr{E}(t)$, and with $\mathbb{P}^*$-probability at least $1 - e^{-t}$ conditioned on $\mathscr{E}(t)$, the bootstrap estimator $\widehat{\boldsymbol{\beta}}^\flat = \widehat{\boldsymbol{\beta}}^\flat(\tau)$ ($0 < \tau < 1$) given in* (2.8) *satisfies*

$$\|\widehat{\boldsymbol{\beta}}^\flat - \boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma}} \lesssim \sqrt{\frac{p+t}{n}} \tag{2.13}$$

as long as $n \gtrsim p + t$.

Theorem 2.1.5. *Suppose that the conditions in Theorem 2.1.2 hold. Under the scaling $n \gtrsim p + \log n$, there exists some event $\mathscr{E}_n$ with $\mathbb{P}(\mathscr{E}_n) \geq 1 - 4n^{-1}$ such that, with $\mathbb{P}^*$-probability at least $1 - n^{-1}$ conditioned on $\mathscr{E}_n$,*

$$\mathbf{S}^{1/2}(\widehat{\boldsymbol{\beta}}^\flat - \widehat{\boldsymbol{\beta}}) = \mathbf{S}^{-1/2}\frac{1}{n}\sum_{i=1}^{n} e_i \boldsymbol{x}_i \big\{\tau - I(\varepsilon_i \leq 0)\big\} + \boldsymbol{r}_n^\flat, \tag{2.14}$$

where $\boldsymbol{r}_n^\flat = \boldsymbol{r}_n^\flat(\{(e_i, y_i, \boldsymbol{x}_i)\}_{i=1}^n)$ satisfies $\|\boldsymbol{r}_n^\flat\|_2 = O_{\mathbb{P}^*}(\chi_n)$, and $\chi_n = \chi_n(\{(y_i, \boldsymbol{x}_i)\}_{i=1}^n)$ is such that $\chi_n = O_{\mathbb{P}}\{(p+\log n)^{1/4}(p\log n)^{1/2} n^{-3/4} + (p+\log n)^{1/2} p \log(n) n^{-1}\}$.

We end this section by validating the (Rademacher) multiplier bootstrap.

Theorem 2.1.6. *Let $\boldsymbol{\lambda} \in \mathbb{R}^p$ be an arbitrary d-vector defining a linear contrast of interest. Assume Conditions 2.1.1 and 2.1.2 hold, and that the parameter dimension p satisfies the scaling*

$p^3(\log n)^2 = o(n)$. *Then, as* $n \to \infty$,

$$\sup_{x\in\mathbb{R}} \left|\mathbb{P}\big(n^{1/2}\langle \boldsymbol{\lambda}, \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\rangle \le x\big) - \mathbb{P}^*\big(n^{1/2}\langle \boldsymbol{\lambda}, \widehat{\boldsymbol{\beta}}^\flat - \widehat{\boldsymbol{\beta}}\rangle \le x\big)\right| \xrightarrow{\mathbb{P}} 0. \tag{2.15}$$

2.1.3 Goodness-of-fit testing

As a by-product, the multiplier bootstrap method can also be applied to goodness-of-fit testing for quantile regression. Under model (2.1), consider a subset $\Omega_0 \subseteq \mathbb{R}_p$ and a test

$$H_0 : \boldsymbol{\beta}^* \in \Omega_0 \quad \text{versus} \quad H_1 : \boldsymbol{\beta}^* \in \mathbb{R}^p \setminus \Omega_0. \tag{2.16}$$

We first construct the test statistics based on the empirical loss $Q_n(\boldsymbol{\beta})$ defined in (2.3). Let $\widehat{\boldsymbol{\beta}}$ be quantile estimator under the full model (2.2), and set $\widehat{\boldsymbol{\beta}}_0 \in \operatorname{argmin}_{\boldsymbol{\beta}\in\Omega_0} Q_n(\boldsymbol{\beta})$. The test statistic is defined as

$$T_n = Q_n(\widehat{\boldsymbol{\beta}}_0) - Q_n(\widehat{\boldsymbol{\beta}}).$$

In the bootstrap world, we intend to mimic the distribution of T_n using that of $Q_n^\flat(\boldsymbol{\beta})$ defined in (2.7). Let $\widehat{\boldsymbol{\beta}}^\flat \in \operatorname{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} Q_n^\flat(\boldsymbol{\beta})$ and $\widehat{\boldsymbol{\beta}}_0^\flat \in \operatorname{argmin}_{\boldsymbol{\beta}\in\Omega_0} Q_n^\flat(\boldsymbol{\beta})$ be the bootstrap statistics in the full model and null model, respectively. Motivated by Chen et al. [2008], we consider the bootstrap test statistic

$$T_n^\flat = \big\{Q_n^\flat(\widehat{\boldsymbol{\beta}}_0^\flat) - Q_n^\flat(\widehat{\boldsymbol{\beta}}^\flat)\big\} - \big\{Q_n^\flat(\widehat{\boldsymbol{\beta}}_0) - Q_n^\flat(\widehat{\boldsymbol{\beta}})\big\}.$$

See Remark 2 therein for the intuition behind this construction. The conditional distribution of $T_n^\flat$ given the data then serves as an approximation of the distribution of T_n. For every $q \in (0,1)$, let γ_q be the (conditional) upper q-quantile of $T_n^\flat$, that is,

$$\gamma_q = \inf\big\{z \in \mathbb{R} : \mathbb{P}^*(T_n^\flat > z) \le q\big\},$$

Consequently, for significance level $\alpha \in (0,1)$, we reject H_0 in (2.16) whenever $T_n > \gamma_\alpha$.

The above method was first proposed and studied by Chen et al. [2008] using standard exponential weights in the case of median regression, and as discussed earlier, the Rademacher multiplier bootstrap is computationally more attractive and also has provable finite-sample guarantees.

2.2 Numerical Experiments

In this section, we conduct numerical experiments to compare the multiplier bootstrap on constructing confidence intervals and goodness-of-fit testing with some well-known existing methods for quantile regression. Our computational results are reproducible using codes available from https://github.com/XiaoouPan/mbQuantile.

2.2.1 Confidence estimation

We first consider the problem of confidence estimation. The limiting distribution of the quantile regression estimator involves the density of the errors, making the non-resampling (plug-in) inference procedure unstable and unreliable. We refer to Kocherginsky, He and Mu [2005] for an overview and numerical comparisons between plug-in and resampling methods. In this paper, we focus on the following bootstrap calibration methods:

- pair: pairwise bootstrap by resampling $\{(y_i, \boldsymbol{x}_i)\}_{i=1}^n$ in pairs with replacement (Section 9.5 of Efron and Tibshirani [1994]);
- pwy: a resampling method based on pivotal estimating functions [Parzen, Wei and Ying, 1994];
- wild: wild bootstrap with Rademacher weights [Feng, He and Hu, 2011];
- mb-per: multiplier bootstrap percentile method defined in (2.10);
- mb-norm: multiplier bootstrap normal-based method defined in (2.11).

The first three methods can be directly implemented using the R package `quantreg`.

To better evaluate the performance of these methods under various environments, we generate data vectors $\{(y_i, \boldsymbol{x}_i)\}_{i=1}^n$ from two types of linear models:

1. (Homoscedastic model):

$$y_i = \beta_0^* + \langle \boldsymbol{x}_i, \boldsymbol{\beta}^* \rangle + \varepsilon_i, \quad i = 1, \ldots, n; \tag{2.17}$$

2. (Heteroscedastic model):

$$y_i = \beta_0^* + \langle \boldsymbol{x}_i, \boldsymbol{\beta}^* \rangle + \frac{2\exp(x_{i1})}{1+\exp(x_{i1})}\varepsilon_i, \quad i = 1, \ldots, n. \tag{2.18}$$

Here we use separate notations to differentiate the intercept β_0^* and coefficient vector $\boldsymbol{\beta}^* \in \mathbb{R}^p$. For each model, we consider three error distributions as follows.

1. t_2: $\varepsilon_i \sim t_2$;

2. Normal mixture type I: $\varepsilon_i = az_1 + (1-a)z_2$, where $a \sim \text{Ber}(0.5)$, $z_1 \sim \mathcal{N}(-1, 1)$ and $z_2 \sim \mathcal{N}(1, 1)$;

3. Normal mixture type II: $\varepsilon_i = az_1 + (1-a)z_2$, where $a \sim \text{Ber}(0.9)$, $z_1 \sim \mathcal{N}(0, 1)$ and $z_2 \sim \mathcal{N}(0, 5^2)$.

Moreover, we generate random predictors with three different covariance structures:

1. Independent design: $\boldsymbol{x}_i \sim \mathcal{N}(0, \mathbf{I}_p)$ for $i = 1, \ldots, n$;

2. Weakly correlated design: first generate a covariance matrix $\boldsymbol{\Sigma} = (\sigma_{jk})_{1 \le j,k \le p}$ with diagonal entries σ_{jj} independently drawn from $\text{Unif}(0.5, 1)$ and $\sigma_{jk} = 0.5^{|j-k|}(\sigma_{jj}\sigma_{kk})^{1/2}$ if $j \neq k$, and then generate $\boldsymbol{x}_i$'s independently from $\mathcal{N}(0, \boldsymbol{\Sigma})$;

3. Equally correlated design: first generate a covariance matrix $\boldsymbol{\Sigma} = (\sigma_{jk})_{1\le j,k\le p}$ with diagonal entries σ_{jj} independently drawn from Unif$(0.5, 1)$ and $\sigma_{jk} = 0.5(\sigma_{jj}\sigma_{kk})^{1/2}$ if $j \neq k$, and then generate $\boldsymbol{x}_i$'s independently from $\mathcal{N}(0, \boldsymbol{\Sigma})$.

We set $\beta_0^* = 2$, $\boldsymbol{\beta}^* = (2, \ldots, 2)^{\mathrm{T}}$ and $(n, p) = (200, 10)$. The confidence level is taken to be $1 - \alpha \in \{80\%, 90\%, 95\%\}$. All of the five methods are carried out using $B = 1000$ bootstrap samples. Tables 2.1 and 2.2 display the average coverage probabilities and average interval widths over all the regression coefficients based on 200 Monte Carlo simulations. We refer to Pan and Zhou [2021] for more comprehensive simulation results.

Table 2.1. Average coverage probabilities and confidence interval (CI) widths over all the coefficients under homoscedastic model (2.17) with type I mixture normal error.

Independent Gaussian design										
	Coverage probability					Width				
α	pair	pwy	wild	mb-per	mb-norm	pair	pwy	wild	mb-per	mb-norm
0.05:	0.963	0.966	0.930	0.967	0.935	0.620	0.635	0.554	0.542	0.540
0.1:	0.922	0.930	0.873	0.925	0.873	0.520	0.533	0.465	0.451	0.453
0.2:	0.828	0.844	0.776	0.824	0.769	0.405	0.415	0.362	0.347	0.353
Weakly correlated Gaussian design										
	Coverage probability					Width				
α	pair	pwy	wild	mb-per	mb-norm	pair	pwy	wild	mb-per	mb-norm
0.05:	0.962	0.966	0.921	0.964	0.926	0.920	0.941	0.815	0.806	0.802
0.1:	0.915	0.921	0.867	0.917	0.873	0.772	0.790	0.684	0.670	0.673
0.2:	0.821	0.835	0.769	0.821	0.767	0.601	0.615	0.533	0.515	0.525
Equally correlated Gaussian design										
	Coverage probability					Width				
α	pair	pwy	wild	mb-per	mb-norm	pair	pwy	wild	mb-per	mb-norm
0.05:	0.964	0.968	0.925	0.967	0.930	0.980	1.004	0.868	0.860	0.856
0.1:	0.913	0.926	0.861	0.921	0.867	0.823	0.842	0.729	0.714	0.718
0.2:	0.826	0.831	0.766	0.816	0.767	0.641	0.656	0.568	0.550	0.559

From Tables 2.1 and 2.2, we find that all the bootstrap methods preserve nominal levels, while pairwise bootstrap and bootstrap based on estimating functions (pwy) tend to be more conservative with wider intervals, and wild bootstrap loses coverage probability under some cases; see Table 2.1. Across all the settings, the multiplier bootstrap methods (percentile and normal-based) provide desirable results in terms of both accuracy (narrow width) and reliability (high confidence). It is worth noticing that the normal-based confidence interval (mb-norm) tends to have lower coverage probabilities compared with the percentile method. As the sample size increases, the coverage probability of mb-norm approaches the nominal level gradually; see

Table 2.2. Average coverage probabilities and CI widths over all the coefficients under heteroscedastic model (2.18) with type I mixture normal error.

Independent Gaussian design										
	Coverage probability					Width				
α	pair	pwy	wild	mb-per	mb-norm	pair	pwy	wild	mb-per	mb-norm
0.05:	0.972	0.974	0.946	0.966	0.945	0.542	0.555	0.481	0.478	0.474
0.1:	0.936	0.938	0.898	0.920	0.905	0.454	0.466	0.404	0.395	0.398
0.2:	0.861	0.870	0.811	0.828	0.805	0.354	0.363	0.315	0.303	0.310
Weakly correlated Gaussian design										
	Coverage probability					Width				
α	pair	pwy	wild	mb-per	mb-norm	pair	pwy	wild	mb-per	mb-norm
0.05:	0.968	0.970	0.941	0.966	0.938	0.820	0.840	0.729	0.722	0.716
0.1:	0.932	0.933	0.885	0.913	0.886	0.688	0.705	0.612	0.597	0.601
0.2:	0.849	0.859	0.791	0.816	0.785	0.536	0.549	0.476	0.458	0.468
Equally correlated Gaussian design										
	Coverage probability					Width				
α	pair	pwy	wild	mb-per	mb-norm	pair	pwy	wild	mb-per	mb-norm
0.05:	0.968	0.974	0.938	0.964	0.941	0.877	0.898	0.778	0.772	0.765
0.1:	0.928	0.932	0.881	0.917	0.883	0.736	0.754	0.653	0.638	0.642
0.2:	0.839	0.847	0.787	0.804	0.786	0.573	0.587	0.509	0.490	0.500

Table 2.3. After taking into account the interval width, we recommend the multiplier bootstrap percentile method that has the best overall performance.

Table 2.3. Average coverage probabilities and CI widths (in brackets) over all the coefficients under homoscedastic model (2.17) with type I mixture normal error.

Independent Gaussian design						
	$n = 200$		$n = 500$		$n = 1000$	
α	mb-per	mb-norm	mb-per	mb-norm	mb-per	mb-norm
0.05:	0.967 (0.542)	0.935 (0.540)	0.950 (0.346)	0.923 (0.346)	0.960 (0.247)	0.948 (0.247)
0.1:	0.925 (0.451)	0.873 (0.453)	0.904 (0.289)	0.871 (0.290)	0.923 (0.206)	0.895 (0.207)
0.2:	0.824 (0.347)	0.769 (0.353)	0.817 (0.224)	0.768 (0.226)	0.824 (0.160)	0.792 (0.161)
Weakly correlated Gaussian design						
	$n = 200$		$n = 500$		$n = 1000$	
α	mb-per	mb-norm	mb-per	mb-norm	mb-per	mb-norm
0.05:	0.964 (0.806)	0.926 (0.802)	0.954 (0.512)	0.933 (0.512)	0.966 (0.364)	0.948 (0.364)
0.1:	0.917 (0.670)	0.873 (0.673)	0.905 (0.428)	0.875 (0.430)	0.913 (0.305)	0.899 (0.306)
0.2:	0.821 (0.515)	0.767 (0.525)	0.798 (0.331)	0.770 (0.335)	0.824 (0.236)	0.799 (0.238)
Equally correlated Gaussian design						
	$n = 200$		$n = 500$		$n = 1000$	
α	mb-per	mb-norm	mb-per	mb-norm	mb-per	mb-norm
0.05:	0.967 (0.860)	0.930 (0.856)	0.960 (0.547)	0.941 (0.546)	0.961 (0.389)	0.944 (0.389)
0.1:	0.921 (0.714)	0.867 (0.718)	0.912 (0.456)	0.873 (0.458)	0.909 (0.326)	0.888 (0.327)
0.2:	0.816 (0.550)	0.767 (0.559)	0.804 (0.353)	0.773 (0.357)	0.818 (0.253)	0.792 (0.255)

Regarding computational complexity, for each bootstrap sample, pairwise and wild bootstraps solve a quantile regression on a sample of size n, bootstrap based on estimating functions (pwy) solves a quantile regression of size $n+1$, while multiplier bootstrap solves a quantile regression essentially on a subsample of size $n/2$ on average. In summary, the multiplier bootstrap provides a computationally efficient way to construct confidence intervals with high

precision and reliability.

2.2.2 Goodness-of-fit testing

In this section, we compare the multiplier bootstrap with classical non-resampling methods on goodness-of-fit testing for quantile regression. Specifically, we consider the following methods:

- Wald: Wald test based on unrestricted estimator [Koenker and Bassett, 1982];
- rank: rank score test [Gutenbrunner et al., 1993];
- mb-exp: multiplier bootstrap with exponential weights [Chen et al., 2008];
- mb-Rad: multiplier bootstrap with Rademacher weights.

The first three methods are included in the R package `quantreg`.

We generate data vectors the same way as in Section 2.2.1. Moreover, we set $(n, p) = (200, 15)$, and the confidence level is taken to be $1-\alpha \in \{90\%, 95\%, 99\%\}$. We consider testing

$$H_0 : \beta_j^* = 0, \text{ for } j = 1, \ldots, 15 \quad \text{versus} \quad H_1 : \beta_j^* \neq 0, \text{ for some } j.$$

To assess the overall performance, we employ the following three measurements:

1. Type I error under null model: $\boldsymbol{\beta}^* = \mathbf{0}$;
2. Power under sparse and strong signal: $\beta_1^* = 0.5$, and $\beta_j^* = 0$ for $j = 2, 3, \ldots, 15$;
3. Power under dense and weak signal: $\beta_j^* = 0.1$ for $j = 1, 2, \ldots, 10$, and $\beta_j^* = 0$ for $j = 11, 12, \ldots, 15$.

The two resampling methods (mb-exp and mb-Rad) are carried out using $B = 1000$ bootstrap samples. Tables 2.4 and 2.5 display the average type I error and power over 200 Monte Carlo simulations. Additional simulation results can be found in Pan and Zhou [2021].

Table 2.4. Average type I error and power under homoscedastic model (2.17) with type I mixture normal error.

Independent Gaussian design												
	Type I error under null model				Power under sparse model				Power under dense model			
α	Wald	rank	mb-exp	mb-Rad	Wald	rank	mb-exp	mb-Rad	Wald	rank	mb-exp	mb-Rad
0.01	0.370	0.000	0.000	0.005	0.805	0.185	0.295	0.330	0.580	0.035	0.045	0.075
0.05	0.490	0.025	0.055	0.050	0.915	0.460	0.570	0.540	0.725	0.150	0.315	0.300
0.1	0.615	0.080	0.140	0.125	0.945	0.625	0.750	0.695	0.775	0.290	0.390	0.360
Weakly correlated Gaussian design												
	Type I error under null model				Power under sparse model				Power under dense model			
α	Wald	rank	mb-exp	mb-Rad	Wald	rank	mb-exp	mb-Rad	Wald	rank	mb-exp	mb-Rad
0.01	0.300	0.010	0.005	0.010	0.650	0.115	0.230	0.250	0.710	0.160	0.210	0.230
0.05	0.450	0.060	0.060	0.055	0.790	0.350	0.465	0.435	0.820	0.380	0.500	0.485
0.1	0.555	0.095	0.120	0.090	0.850	0.515	0.605	0.575	0.870	0.535	0.640	0.600
Equally correlated Gaussian design												
	Type I error under null model				Power under sparse model				Power under dense model			
α	Wald	rank	mb-exp	mb-Rad	Wald	rank	mb-exp	mb-Rad	Wald	rank	mb-exp	mb-Rad
0.01	0.300	0.010	0.010	0.010	0.660	0.135	0.205	0.225	0.915	0.470	0.595	0.615
0.05	0.450	0.060	0.060	0.055	0.790	0.325	0.400	0.385	0.960	0.755	0.825	0.800
0.1	0.555	0.095	0.120	0.090	0.870	0.460	0.575	0.515	0.970	0.860	0.860	0.860

Table 2.5. Average type I error and power under heteroscedastic model (2.18) with type I mixture normal error.

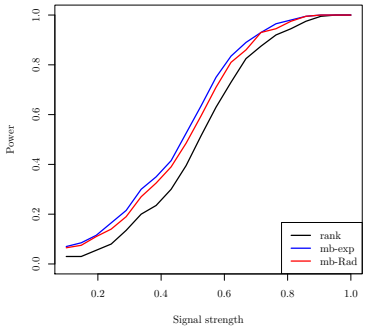
Independent Gaussian design												
	Type I error under null model				Power under sparse model				Power under dense model			
α	Wald	rank	mb-exp	mb-Rad	Wald	rank	mb-exp	mb-Rad	Wald	rank	mb-exp	mb-Rad
0.01	0.315	0.005	0.000	0.000	0.815	0.410	0.475	0.510	0.590	0.085	0.095	0.110
0.05	0.435	0.035	0.030	0.030	0.930	0.685	0.755	0.725	0.705	0.275	0.305	0.305
0.1	0.510	0.065	0.065	0.050	0.955	0.785	0.840	0.810	0.780	0.415	0.415	0.380
Weakly correlated Gaussian design												
	Type I error under null model				Power under sparse model				Power under dense model			
α	Wald	rank	mb-exp	mb-Rad	Wald	rank	mb-exp	mb-Rad	Wald	rank	mb-exp	mb-Rad
0.01	0.380	0.010	0.005	0.005	0.810	0.200	0.330	0.365	0.790	0.235	0.260	0.295
0.05	0.480	0.060	0.055	0.050	0.885	0.525	0.610	0.565	0.865	0.510	0.595	0.565
0.1	0.565	0.110	0.115	0.090	0.905	0.655	0.740	0.700	0.910	0.680	0.725	0.690
Equally correlated Gaussian design												
	Type I error under null model				Power under sparse model				Power under dense model			
α	Wald	rank	mb-exp	mb-Rad	Wald	rank	mb-exp	mb-Rad	Wald	rank	mb-exp	mb-Rad
0.01	0.350	0.010	0.005	0.005	0.690	0.205	0.270	0.300	0.960	0.610	0.715	0.735
0.05	0.470	0.060	0.045	0.040	0.815	0.450	0.550	0.520	0.990	0.850	0.900	0.880
0.1	0.535	0.125	0.115	0.100	0.865	0.610	0.695	0.640	0.990	0.900	0.935	0.935

From Tables 2.4 and 2.5 we see that the Wald test suffers from severe size distortion by rejecting much more often than it should, while the other three methods have type I errors close to the nominal level. Under both sparse and dense alternatives, the multiplier bootstrap outperforms the rank score test with higher power throughout all the combinations of design and error distributions.

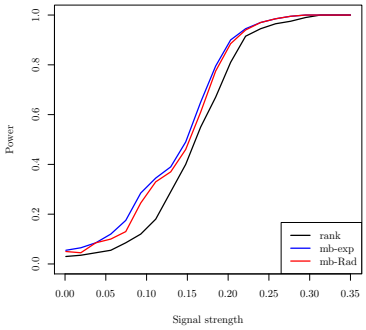
To further compare the power of the last three methods, we draw the power curve with gradually increasing signal strength under sparse and dense settings. Figure 2.1 is a visualization of Table 2.4 and Table 2.5 with type I mixture normal error and independent design. The advantage of multiplier bootstrap over rank test is conspicuous under homoscedastic model, and multiplier bootstrap reveals perceptible advantage as signal gets stronger under heteroscedastic model.

2.3 Acknowledgements

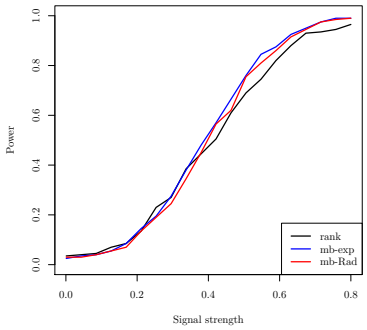
This chapter, in part, is a reprint of the material in the paper "Multiplier bootstrap for quantile regression: Non-asymptotic theory under random design", Pan, Xiaoou and Zhou, Wen-Xin. The paper has been published on *Information and Inference: A Journal of the IMA*, **10** 813-861. The dissertation author was the primary investigator and author of this paper.



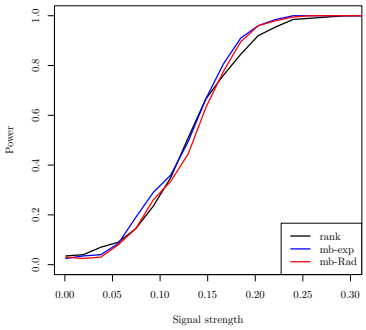
(a) Homoscedastic model (2.17) with sparse signal.



(b) Homoscedastic model (2.17) with dense signal.



(c) Heteroscedastic model (2.18) with sparse signal.



(d) Heteroscedastic model (2.18) with dense signal.

Figure 2.1. Power curves of the three methods under independent design and type I mixture normal error with $\alpha = 0.05$.

Chapter 3

Scalable Learning via Convolution-type Smoothing

3.1 Smoothed Quantile Regression

3.1.1 Motivation and overview

Recall that given a random sample $\{(y_i, \boldsymbol{x}_i)\}_{i=1}^n$, the standard quantile regression estimator is obtained as

$$\widehat{\boldsymbol{\beta}}(\tau) \in \min_{\boldsymbol{\beta}\in\mathbb{R}^p} \widehat{Q}(\boldsymbol{\beta}) = \min_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^n \rho_\tau(y_i - \langle \boldsymbol{x}_i, \boldsymbol{\beta}\rangle), \tag{3.1}$$

where $\rho_\tau(u) = u\{\tau - \mathbb{1}(u < 0)\}$. To circumvent the non-differentiability stemmed from the indicator of the QR loss function, Horowitz [1998] proposed to smooth the indicator part of the check function via the survival function of a kernel. This smoothing method, which we refer to as *Horowitz's smoothing* throughout, has been widely used for various QR-related problems with complex data [Wang, Stefanski and Zhu, 2012, Wu, Ma and Yin, 2015, Galvao and Kato, 2016, de Castro et al., 2019, Chen, Liu and Zhang, 2019]. However, Horowitz's smoothing gains smoothness at the cost of convexity, which inevitably raises optimization-related issues. In general, computing a global minimum of a non-convex function is intractable: finding an ε-suboptimal point for a k-times continuously differentiable function $f : \mathbb{R}^p \to \mathbb{R}$ requires at least as many as $(1/\varepsilon)^{p/k}$ evaluations of the function and its first k derivatives [Nemirovski and

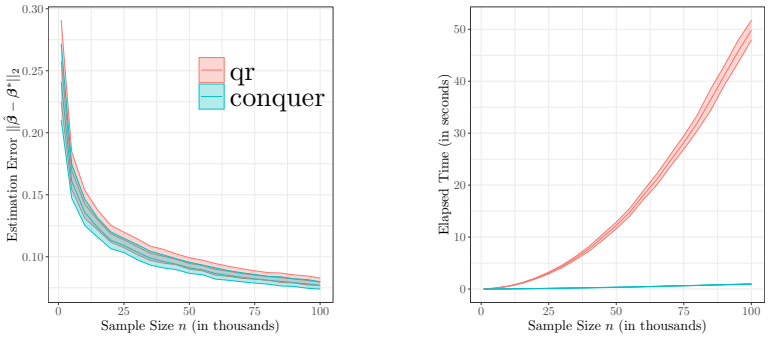
Yudin, 1983]. As we shall see from the numerical studies in Section 3.4, the convergence of gradient-based algorithms can be relatively slow for high and low quantile levels. To address the aforementioned issue, Fernandes, Guerre and Horta [2021] proposed a convolution-type smoothing method that yields a convex and twice differentiable loss function, and studied the asymptotic properties of the smoothed estimator when p is fixed. To distinguish this approach from Horowitz's smoothing, we adopt the term *conquer* for convolution-type smoothed quantile regression.

In this chapter, we first provide an in-depth statistical analysis of conquer under various nonstandard asymptotics settings in which p increases with n. Our results reveal a key feature of the smoothing parameter, often referred to as the bandwidth: the bandwidth adapts to both the sample size n and dimensionality p, so as to achieve a tradeoff between statistical accuracy and computational stability. Since the convolution smoothed loss function is globally convex and locally strongly convex, we propose an efficient gradient descent algorithm with the Barzilai-Borwein stepsize and a Huber-type initialization. The proposed algorithm is implemented via `RcppArmadillo` [Eddelbuettel and Sanderson, 2014] in the `R` package `conquer`. We next focus on large-scale statistical inference (hypothesis testing and confidence estimation) with large p and larger n. We propose a bootstrapped conquer method that has reduced computational complexity when the conquer estimator is used as initialization. Under appropriate restrictions on dimension, we establish the consistency (or concentration), Bahadur representation, asymptotic normality of the conquer estimator as well as the validity of the bootstrap approximation. In the following, we provide more details on the computational and statistical contributions of this paper.

Theoretically, by allowing p to grow with n, the "complexity" of the function classes that we come across in the analysis also increases with n. Conventional asymptotic tools for proving the bootstrap validity are based on weak convergence arguments [van der Vaart and Wellner, 1996], which are not directly applicable in the increasing dimension setting, especially with a non-differentiable loss. In this paper we turn to a more refined and self-contained analysis,

and prove a new local restricted strong convexity (RSC) property for the empirical smoothed quantile loss. This validates the key merit of convolution-type smoothing, i.e., local strong convexity. The smoothing method involves a bandwidth, denoted by h. Theoretically, we show that with sub-exponential random covariates (relaxing the bounded covariates assumption in Fernandes, Guerre and Horta [2021]), conquer exhibits an ℓ_2-error of order $\sqrt{(p+t)/n}+h^2$ with probability at least $1-2e^{-t}$. When h is of order $\{(p+\log n)/n\}^\gamma$ for any $\gamma \in [1/4, 1/2]$, the conquer estimation is first-order equivalent to QR. Under slightly more stringent sub-Gaussian condition on the covariates, we show that the Bahadur-Kiefer linearization error of conquer is of order $(p+t)/(nh^{1/2})+h^{3/2}\sqrt{(p+t)/n}+h^4$ with probability at least $1-3e^{-t}$. Based on such a representation, we establish a Berry-Esseen bound for linear functionals of conquer, which lays the theoretical foundation for testing general linear hypotheses, encompassing covariate-effect analysis, analysis of variance, and model comparisons, to name a few. It is worth noting that with a properly chosen h, the linear functional of conquer is asymptotically normal as long as $p^{8/3}/n \to 0$, which improves the best known growth condition on p for standard QR [Welsh, 1989, He and Shao, 2000, Pan and Zhou, 2021]. We attribute this gain to the effect of smoothing. Under similar conditions, we further establish upper bounds on both estimation and Bahadur-Kiefer linearization errors for the bootstrapped conquer estimator.

To better appreciate the computational feasibility of conquer for large-scale problems, we compare it with standard QR on large synthetic datasets, where the latter is implemented by the `R` package `quantreg` [Koenker, 2022] using the Frisch-Newton approach after preprocessing "pfn". We generate independent data vectors $\{y_i, \boldsymbol{x}_i\}_{i=1}^n$ from a linear model $y_i = \beta_0^* + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^* + \varepsilon_i$, where $(\beta_0^*, \boldsymbol{\beta}^{*\mathrm{T}})^{\mathrm{T}} = (1, \ldots, 1)^{\mathrm{T}} \in \mathbb{R}^{p+1}$, $\boldsymbol{x}_i \sim \mathscr{N}_p(0, \mathbf{I})$ and $\varepsilon_i \sim t_2$. We report the estimation error and elapsed time for increasing sample sizes $n \in \{1000, 5000, 10000, \ldots, 100000\}$ and dimension $p = \lfloor n^{1/2} \rfloor$, the largest integer that is less than or equal to $n^{1/2}$. Figure 3.1 displays the average estimation error, average elapsed time and their standard deviations based on 100 Monte Carlo samples. This experiment shows promise of conquer as a practically useful tool for large-scale quantile regression analysis. More empirical evidence will be given in the latter section.



(a) Estimation error **(b)** Runtime

Figure 3.1. A numerical comparison between conquer and QR. The latter is implemented by the `R` package `quantreg` using the "pfn" method. Panels (a) and (b) display, respectively, the "estimation error and its standard deviation versus sample size" and "elapsed time and its standard deviation versus sample size" as the size of the problem increases.

3.1.2 Convolution-type smoothing

Let $Q(\boldsymbol{\beta}) = \mathbb{E}\{\widehat{Q}(\boldsymbol{\beta})\}$ be the population quantile loss function. Under mild conditions, $Q(\cdot)$ is twice differentiable and strongly convex in a neighborhood of $\boldsymbol{\beta}^*$ with Hessian matrix $\mathbf{J} := \nabla^2 Q(\boldsymbol{\beta}^*) = \mathbb{E}\{f_{\varepsilon|\boldsymbol{x}}(0)\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\}$, where $\varepsilon = y - \langle \boldsymbol{x}, \boldsymbol{\beta}^*(\tau)\rangle$ is the random noise and $f_{\varepsilon|\boldsymbol{x}}(\cdot)$ is the conditional density of ε given $\boldsymbol{x}$. In contrast, the empirical quantile loss $\widehat{Q}(\cdot)$ is not differentiable at $\boldsymbol{\beta}^*$, and its "curvature energy" is concentrated at a single point. This is substantially different from other widely used loss functions that are at least locally strongly convex, such as the squared or logistic loss. The non-smoothness property not only brings challenge to theoretical analysis, but more importantly, also prevents gradient-based optimization methods from being efficient. In his seminal work, Horowitz [1998] proposed to directly smooth the check function $\rho_\tau(\cdot)$ to obtain

$$\ell_h^{\mathrm{Horo}}(u) = u\big\{\tau - \mathscr{G}(-u/h)\big\}, \tag{3.2}$$

where $\mathscr{G}(\cdot)$ is a smooth function that takes values between 0 and 1, and $h > 0$ is a smoothing parameter/bandwidth. However, Horowitz's smoothing gains smoothness at the cost of convexity,

which inevitably raises optimization issues especially when p is large. On the other hand, by the first-order condition, the population parameter $\boldsymbol{\beta}^*$ satisfies the moment condition

$$\nabla Q(\boldsymbol{\beta}^*) = \mathbb{E}\left[\left\{\mathbb{1}(y < \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}) - \tau\right\}\boldsymbol{x}\right]\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*} = \mathbf{0}.$$

This property motivates a smoothed estimating equation (SEE) estimator [Whang, 2006, Kaplan and Sun, 2017], defined as the solution to the smoothed moment condition

$$\frac{1}{n}\sum_{i=1}^{n}\left[\mathscr{G}\left\{(\langle \boldsymbol{x}_i, \boldsymbol{\beta}\rangle - y_i)/h\right\} - \tau\right]\boldsymbol{x}_i = \mathbf{0}. \tag{3.3}$$

Let $K(\cdot)$ be a kernel function that integrates to one, and $h > 0$ be a bandwidth. Throughout the paper, we write

$$K_h(u) = h^{-1}K(u/h), \quad \mathscr{K}_h(u) = \mathscr{K}(u/h) \text{ and } \mathscr{K}(u) = \int_{-\infty}^{u} K(v)\,\mathrm{d}v, \ \ u \in \mathbb{R}. \tag{3.4}$$

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

$$\widehat{Q}_h(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\ell_h(y_i - \langle \boldsymbol{x}_i, \boldsymbol{\beta}\rangle) \quad \text{with} \quad \ell_h(u) = (\rho_\tau * K_h)(u) = \int_{-\infty}^{\infty}\rho_\tau(v)K_h(v-u)\,\mathrm{d}v, \tag{3.5}$$

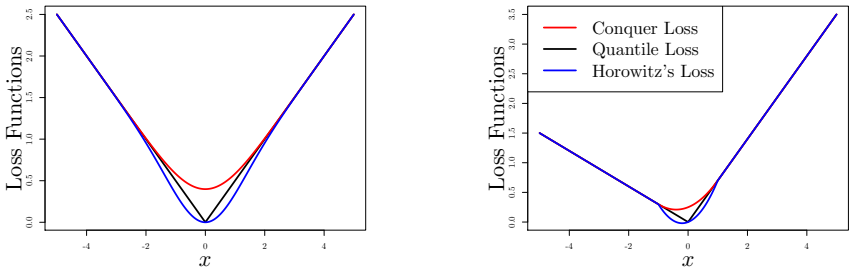
where $*$ denotes the convolution operator. Therefore, as stated in the Introduction, we refer to the aforementioned smoothing method as *conquer*. The ensuing conquer estimator is given by

$$\widehat{\boldsymbol{\beta}}_h = \widehat{\boldsymbol{\beta}}_h(\tau) \in \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p}\widehat{Q}_h(\boldsymbol{\beta}). \tag{3.6}$$

The key difference between the conquer loss (3.5) and Horowitz's loss (3.2) is that the former is globally convex, while Horowitz's loss is not. This is illustrated in Figure 3.2.

As we shall see later, the ideal choice of bandwidth should adapt to the sample size n and

dimension p, since the quantile level τ is prespecified and fixed. Thus, the dependence of $\widehat{\boldsymbol{\beta}}_h$ and $\widehat{Q}_h(\cdot)$ on τ will be assumed without display. Commonly used kernel functions include: (a) uniform kernel $K(u) = (1/2)\mathbb{1}(|u| \leq 1)$, (b) Gaussian kernel $K(u) = \phi(u) := (2\pi)^{-1/2} e^{-u^2/2}$, (c) logistic kernel $K(u) = e^{-u}/(1+e^{-u})^2$, (d) Epanechnikov kernel $K(u) = (3/4)(1-u^2)\mathbb{1}(|u| \leq 1)$, and (e) triangular kernel $K(u) = (1-|u|)\mathbb{1}(|u| \leq 1)$. Explicit expressions of the corresponding smoothed loss function $\rho_\tau * K_h$ will be given in Section 3.2.



(a) Guassian kernel under $\tau = 0.5$. **(b)** Uniform kernel under $\tau = 0.7$.

Figure 3.2. Quantile loss in (3.1), conquer loss (3.5), and Horowitz's smoothed loss (3.2) with Gaussian and uniform kernels, respectively.

The convolution-type kernel smoothing yields an objective function $\boldsymbol{\beta} \mapsto \widehat{Q}_h(\boldsymbol{\beta})$ that is twice continuously differentiable with gradient and hessian matrix

$$\nabla \widehat{Q}_h(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\big\{\mathscr{K}_h\big(\langle \boldsymbol{x}_i, \boldsymbol{\beta}\rangle - y_i\big) - \tau\big\}\boldsymbol{x}_i \quad \text{and} \quad \nabla^2 \widehat{Q}_h(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} K_h(y_i - \langle \boldsymbol{x}_i, \boldsymbol{\beta}\rangle)\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}, \quad (3.7)$$

respectively, where $\mathscr{K}_h(\cdot) = \mathscr{K}(\cdot/h)$ is defined in (3.4). Provided that K is non-negative, $\widehat{Q}_h(\cdot)$ is a convex function for any $h > 0$, and $\widehat{\boldsymbol{\beta}}_h = \widehat{\boldsymbol{\beta}}_h(\tau)$ satisfies the first-order condition $\nabla \widehat{Q}_h(\widehat{\boldsymbol{\beta}}_h) = \mathbf{0}$. This reveals the connection between the SEE and the conquer methods. Together, the smoothness and convexity of $\widehat{Q}_h(\cdot)$ warrant the superior computation efficiency of first-order gradient based algorithms for solving large-scale smoothed quantile regressions. The computational aspect of conquer will be discussed in Section 3.2.

When the dimension p is fixed, asymptotic properties of the SEE or conquer estimator

have been studied by Kaplan and Sun [2017] and Fernandes, Guerre and Horta [2021]. The former used a higher-order kernel to deal with the instrumental variables QR problem (see Section 3.1.4 for further discussions), and the latter showed that the conquer estimator has a lower asymptotic mean squared error than Horowitz's smoothed estimator, and also has a smaller Bahadur linearization error than the standard QR in the almost sure sense. The optimal order of the bandwidth based on the asymptotic mean squared error is unveiled as a function of n. In Section 3.3, we will establish exponential concentration inequalities and non-asymptotic Bahadur representation for the conquer estimator, while allowing the dimension p to grow with the sample size n. Our results reveal a key feature of the smoothing parameter: the bandwidth should adapt to both the sample size n and dimensionality p, so as to achieve a tradeoff between statistical accuracy and computational stability.

Remark 3.1.1. *As discussed in Fernandes, Guerre and Horta [2021], another advantage of convolution smoothing is that it facilitates conditional density estimation for the quantile regression process. Assume* $Q_y(\tau|\boldsymbol{x}) = F_{y|\boldsymbol{x}}^{-1}(\tau) = \langle \boldsymbol{x}, \boldsymbol{\beta}^*(\tau)\rangle$ *for all* $\tau \in [\tau_L, \tau_U] \subseteq (0,1)$. *Under mild regularity conditions,* $q_y(\tau|\boldsymbol{x}) := \partial Q_y(\tau|\boldsymbol{x})/\partial\tau = 1/f_{y|\boldsymbol{x}}(\langle \boldsymbol{x}, \boldsymbol{\beta}^*(\tau)\rangle)$ *exists. The inverse conditional density function plays an important role in, for example, the study of quantile treatment effects through modeling inverse propensity scores [Firpo, 2007, Chen, Hong and Tarozzi, 2008]. By the linear conditional quantile model assumption,* $\partial Q_y(\tau|\boldsymbol{x})/\partial\tau = \langle \boldsymbol{x}, \partial\boldsymbol{\beta}^*(\tau)/\partial\tau\rangle$ *for* $\tau \in (\tau_L, \tau_U)$. *Recall that the conquer estimator* $\widehat{\boldsymbol{\beta}}_h = \widehat{\boldsymbol{\beta}}_h(\tau)$ *satisfies the first-order condition* $\nabla\widehat{Q}_h(\widehat{\boldsymbol{\beta}}_h(\tau)) = 0$. *Taking the partial derivative with respect to* τ *on both sides, it follows from* (3.7) *and the chain rule that*

$$\frac{\partial\widehat{\boldsymbol{\beta}}_h(\tau)}{\partial\tau} = \big\{\nabla^2\widehat{Q}_h\big(\widehat{\boldsymbol{\beta}}_h(\tau)\big)\big\}^{-1}\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i = \left\{\frac{1}{n}\sum_{i=1}^n K_h\big(y_i - \boldsymbol{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_h(\tau)\big)\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}\right\}^{-1}\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i.$$

Consequently, the inverse densities $1/f_{y_i|\boldsymbol{x}_i}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*(\tau))$ *can be directly estimated by* $\boldsymbol{x}_i^{\mathrm{T}}\frac{\partial\widehat{\boldsymbol{\beta}}_h(\tau)}{\partial\tau}$. *This bypasses the use of any nonparametric method for density estimation with fitted residuals.*

3.1.3 Multiplier bootstrap inference

In this section, we apply the multiplier bootstrap procedure discussed in Section 2.1.2 to construct confidence intervals for conquer. Specifically, define the weighted quantile loss $\widehat{Q}_h^\flat : \mathbb{R}^p \to \mathbb{R}$ as

$$\widehat{Q}_h^\flat(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} w_i \ell_h(y_i - \langle \boldsymbol{x}_i, \boldsymbol{\beta}\rangle), \tag{3.8}$$

where $\ell_h(u) = (\rho_\tau * K_h)(u)$ is as in (3.5) and $w_i \in \{0, 2\}$. The ensuing multiplier bootstrap statistic is then

$$\widehat{\boldsymbol{\beta}}_h^\flat = \widehat{\boldsymbol{\beta}}_h^\flat(\tau) \in \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} \widehat{Q}_h^\flat(\boldsymbol{\beta}). \tag{3.9}$$

Consequently, element-wise confidence intervals can be constructed using one of the three methods mentioned in Section 2.1.2.

In the next section, we will present a finite-sample theoretical framework for convolution-type smoothed quantile regression, including the concentration inequality and non-asymptotic Bahadur representation for both the conquer estimator (3.6) and its bootstrap counterpart (3.9) using Rademacher multipliers. As a by-product, a Berry-Esseen-type inequality (see Theorem 2.1.3) states that, under certain constraints on the (growing) dimensionality and bandwidth, the distribution of any linear projection of $\widehat{\boldsymbol{\beta}}_h$ converges to a normal distribution as the sample size increases to infinity. Informally, for any given deterministic vector $\boldsymbol{a} \in \mathbb{R}^p$, the scaled statistic $n^{1/2}\langle \boldsymbol{a}, \widehat{\boldsymbol{\beta}}_h - \boldsymbol{\beta}^*\rangle$ is asymptotically normally distributed with asymptotic variance $\sigma_0^2(\boldsymbol{a}) := \tau(1-\tau)\,\boldsymbol{a}^{\mathrm{T}}\mathbf{J}^{-1}\boldsymbol{\Sigma}\mathbf{J}^{-1}\boldsymbol{a}$, where $\boldsymbol{\Sigma}$ is the population covariance matrix of the covariates $\boldsymbol{x}$. Another interesting implication from our theoretical analysis is that the unit variance requirement $\mathrm{var}(w_i) = 1$ for the random weight is not necessary to ensure (asymptotically) valid bootstrap inference after a proper variance adjustment. See Remark 3.3.3 for more details.

To make inference based on such asymptotic results, we need to consistently estimate the

asymptotic variance. Fernandes, Guerre and Horta [2021] suggested the following estimators

$$\widehat{\mathbf{J}}_h := \nabla^2 \widehat{\mathcal{Q}}_h(\widehat{\boldsymbol{\beta}}_h) = \frac{1}{nh}\sum_{i=1}^{n} K(\widehat{\varepsilon}_i/h)\cdot \boldsymbol{x}\boldsymbol{x}_i^{\mathrm{T}} \quad \text{and} \quad \widehat{\mathbf{V}}_h := \frac{1}{n}\sum_{i=1}^{n}\big\{\mathscr{K}_h(-\widehat{\varepsilon}_i)-\tau\big\}^2 \boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}} \tag{3.10}$$

of $\mathbf{J}$ and $\tau(1-\tau)\boldsymbol{\Sigma}$, respectively, where $\widehat{\varepsilon}_i = y_i - \langle \boldsymbol{x}_i, \widehat{\boldsymbol{\beta}}_h\rangle$ are fitted residuals. The ensuing $1-\alpha$ normal-based CIs are given by $\widehat{\beta}_{h,j} \pm \Phi^{-1}(1-\alpha/2)\cdot n^{-1/2}(\widehat{\mathbf{J}}_h^{-1}\widehat{\mathbf{V}}_h\widehat{\mathbf{J}}_h^{-1})_{jj}^{1/2}$, $j=1,\ldots,p$. The normal approximations to the CI may suffer from the sensitivity to the smoothing needed to estimate the conditional densities, namely, the matrix $\mathbf{J} = \mathbb{E}\{f_{\varepsilon|\boldsymbol{x}}(0)\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\}$. When p is large, inverting the estimated density matrix $\widehat{\mathbf{J}}_h$ may be numerically unstable. This is typically true when τ is in the upper or lower tail.

3.1.4 Connections to instrumental variable quantile regression

This work focuses on large-scale estimation and inference for linear quantile regression with many exogenous covariates. However, in many economic applications, some regressors of interest (e.g., education, prices) are endogenous, making conventional quantile regression inconsistent for estimating causal quantile effects. To address this problem, Chernozhukov and Hansen [2005] proposed an instrumental variable quantile regression (IVQR) model, which has become a popular tool for estimating quantile effects with endogenous covariates. Due to the non-convex and non-smooth nature of the problem, there is a burgeoning literature on estimation and inference of IVQR models and the related computational issues, dating back to Chen, Linton and Van Keilegom [2003] and Chernozhukov and Hansen [2006]. We refer to Chernozhukov, Hansen and Wüthrich [2020]—Chapter 9 of Koenker et al. [2017]—for an overview of IVQR modeling, from identification conditions to estimation and inference. More specifically, see Horowitz and Lee [2007] and Chen and Pouzo [2009] for non- and semi-parametric IVQR estimation; Kaplan and Sun [2017] and de Castro et al. [2019] for smoothed methods; and Chen and Lee [2018] and Zhu [2018] for methods based on reformulation as mixed integer optimization (MIO), Machado and Santos Silva [2018] for moment-based estimators, and Kaido

and Wüthrich [2021] for a decentralization approach which decomposes the IVQR estimation problem into a set of conventional QR sub-problems.

The convolution smoothing method studied in this paper can be directly linked to the SEE approach in Kaplan and Sun [2017]. The latter addressed the more challenging IVQR problem, and derived both asymptotic mean squared error and normality for the SEE estimator when the dimension is fixed. Our study complements that of Kaplan and Sun [2017] in two ways. First, we provide a systematic analysis for smoothed (conventional) QR from an M-estimation viewpoint under the growing dimension setting. Our results provide explicit finite-sample bounds for the estimation error, Bahadur linearization error as long as their (multiplier) bootstrap counterparts. Asymptotic validity of the multiplier bootstrap is also rigorously established. Secondly, we propose tailored computational methods for smoothed QR computation, which rely on the use of non-negative kernels and the resulting local strong convexity. Compared with generic optimization toolboxes for solving linear programs, the computational efficiency of the gradient-based algorithm for conquer is considerably improved, especially for large-scale problems with many (exogenous) regressors and massive sample size. A potential application is empirical asset pricing via quantile regression, extending the existing machine learning tools for average return forecasting [Gu, Kelly and Xiu, 2020].

In the presence of both exogenous and endogenous covariates, the advantage of smoothing is diluted because the non-convexity issue prevails. The MIO-based IVQR estimation procedure can be implemented by the Gurobi commercial MIO solver, which is free for academic use. The MIO solver converges fast when the number of endogeous covariates varies in the range of 5 and 20 [Zhu, 2018]. The MIO solver in moderate dimensions typically takes much longer to complete the optimization: optimal solutions may be found in a few seconds, but it can take much longer to certify optimality via lower bounds [Bertsimas, King and Mazumder, 2016]. [1]

Recently, Kaido and Wüthrich [2021] proposed a "decentralized" approach for IVQR es-

[1]MIO solvers provide both feasible solutions and lower bounds to the optimal value. As the MIO solver progresses toward the optimal solution, the lower bounds improve and provide an increasingly better guarantee of suboptimality. It is the lower bounds that take so long to converge.

timation. The idea is to decompose the non-convex program into $p_d + 1$ conventional (weighted) quantile regression sub-problems. The IVQR estimator is then characterized as a fixed point of such sub-problems. Since p_d—the number of endogenous variables—is typically small, the overall computational complexity depends primarily on the QR fitting step. When the number of exogenous variables, p_x, is large in the range of hundreds to thousands, the proposed framework in this paper, along with the accompanying software `conquer`, provides a viable option to further reduce the computational cost of the above IVQR estimation method. We leave a rigorous theoretical investigation (when p_d is fixed, $p_x = p_x(n) \to \infty$ and $p_x/n \to 0$ as $n \to \infty$) as well as empirical applications with many (exogenous) regressors to future work.

3.2 Computational Methods

To solve optimization problems (3.6) and (3.9) with non-negative weights, arguably the simplest algorithm is a vanilla gradient descent algorithm (GD). For a prespecified $\tau \in (0,1)$ and bandwidth $h > 0$, recall that $\widehat{Q}_h(\boldsymbol{\beta}) = (1/n)\sum_{i=1}^n \ell_h(y_i - \langle \boldsymbol{x}_i, \boldsymbol{\beta} \rangle)$. Starting with an initial value $\boldsymbol{\beta}^0 \in \mathbb{R}^p$, at iteration $t = 0, 1, 2, \ldots$, GD computes

$$\boldsymbol{\beta}^{t+1} = \boldsymbol{\beta}^t - \eta_t \cdot \nabla \widehat{Q}_h(\boldsymbol{\beta}^t) = \boldsymbol{\beta}^t - \frac{\eta_t}{n} \sum_{i=1}^n \big\{ \mathscr{K}_h(\langle \boldsymbol{x}_i, \boldsymbol{\beta}^t \rangle - y_i) - \tau \big\} \boldsymbol{x}_i, \tag{3.11}$$

where $\eta_t > 0$ is the stepsize. In the classical GD method, the stepsize is usually obtained by employing line search techniques. However, line search is computationally intensive for large-scale settings. One of the most important issues in GD is to determine a proper update step η_t decay schedule. A common practice in the literature is to use a diminishing stepsize or a best-tuned fixed stepsize. Neither of these two approaches can be efficient, at least compared to the Newton-Frisch algorithm with preprocessing [Portnoy and Koenker, 1997]. Recall that the smoothed loss $\widehat{Q}_h(\cdot)$ is twice differentiable with Hessian $\nabla^2 \widehat{Q}_h(\boldsymbol{\beta}) = (1/n)\sum_{i=1}^n K_h(y_i - \langle \boldsymbol{x}_i, \boldsymbol{\beta} \rangle) \boldsymbol{x}_i \boldsymbol{x}_i^{\mathrm{T}}$. It is therefore natural to employ the Newton-Raphson method, which at iteration t

would read

$$\boldsymbol{\beta}^{t+1} = \boldsymbol{\beta}^t + \boldsymbol{d}^t \quad \text{with} \quad \boldsymbol{d}^t := -\big\{\nabla^2 \widehat{Q}_h(\boldsymbol{\beta}^t)\big\}^{-1} \nabla \widehat{Q}_h(\boldsymbol{\beta}^t). \tag{3.12}$$

In practice, the Newton method is often paired with Armoji stepsize: choose a stepsize $\lambda^t = \max\{1, 1/2, 1/4, \ldots\}$ such that $\widehat{Q}_h(\boldsymbol{\beta}^t) - \widehat{Q}_h(\boldsymbol{\beta}^t + \lambda^t \boldsymbol{d}^t) \geq -c\lambda^t \nabla \widehat{Q}_h(\boldsymbol{\beta}^t)\boldsymbol{d}^t$, where $c \in (0, 1/2)$. Then redefine the current iterate as $\boldsymbol{\beta}^{t+1} = \boldsymbol{\beta}^t + \lambda^t \boldsymbol{d}^t$. Since such a backtracking line search requires evaluations of the loss function itself, in the following remark we present the explicit expressions of the convolution smoothed check function for several commonly used kernels.

Remark 3.2.1. *Recall that the check function can be written as $\rho_\tau(u) = |u|/2 + (\tau - 1/2)u$, which, after convolution smoothing, becomes $\ell_h(u) = (1/2)\int_{-\infty}^{\infty} |u + hv| K(v)\,\mathrm{d}v + (\tau - 1/2)u$.*

- *(Gaussian kernel $K(u) = (2\pi)^{-1/2} e^{-u^2/2}$): $\ell_h(u) = (h/2)\ell^{\mathrm{G}}(u/h) + (\tau - 1/2)u$, where $\ell^{\mathrm{G}}(u) := (2/\pi)^{1/2} e^{-u^2/2} + u\{1 - 2\Phi(-u)\}$.*

- *(Logistic kernel[2] $K(u) = e^{-u}/(1+e^{-u})^2$): $\ell_h(u) = (h/2)\ell^{\mathrm{L}}(u/h) + (\tau - 1/2)u$, where $\ell^{\mathrm{L}}(u) := u + 2\log(1 + e^{-u})$.*

- *(Uniform kernel $K(u) = (1/2)\mathbb{1}(|u| \leq 1)$): $\ell_h(u) = (h/2)\ell^{\mathrm{U}}(u/h) + (\tau - 1/2)u$, where $\ell^{\mathrm{U}}(u) := (u^2/2 + 1/2)\mathbb{1}(|u| \leq 1) + |u|\mathbb{1}(|u| > 1)$ is a shifted Huber loss [Huber, 1973].*

- *(Epanechnikov kernel $K(u) = (3/4)(1 - u^2)\mathbb{1}(|u| \leq 1)$): $\ell_h(u) = (h/2)\ell^{\mathrm{E}}(u/h) + (\tau - 1/2)u$, where $\ell^{\mathrm{E}}(u) := (3u^2/4 - u^4/8 + 3/8)\mathbb{1}(|u| \leq 1) + |u|\mathbb{1}(|u| > 1)$.*

- *(Triangular kernel $K(u) = (1 - |u|)\mathbb{1}(|u| \leq 1)$): $\ell_h(u) = (h/2)\ell^{\mathrm{T}}(u/h) + (\tau - 1/2)u$, where $\ell^{\mathrm{T}}(u) := (u^2 - |u|^3/3 + 1/3)\mathbb{1}(|u| \leq 1) + |u|\mathbb{1}(|u| > 1)$.*

[2] Logistic kernel smoothed approximation of the check function dates back to Amemiya [1982], which is used as a technical device to simplify the analysis of the asymptotic behavior of a two-stage median regression estimator.

3.2.1 The Barzilai-Borwein stepsize

In this section, we propose to solve conquer by means of the gradient descent with a Barzilai-Borwein update step [Barzilai and Borwein, 1988], which we refer to as the GD-BB algorithm. Motivated by quasi-Newton methods, the BB method has been proven to be very successful in solving nonlinear optimization problems.

Computing the inverse of the Hessian when p is large is an expensive operation at each Newton step (3.12). Moreover, in circumstances where h is small or τ is very close to 0 or 1, $\nabla^2 \widehat{Q}_h(\cdot)$ may have a large condition number, thus leading to slow convergence. For this reason, many quasi-Newton methods seek a simple approximation of the inverse Hessian matrix, say $(\mathbf{J}^t)^{-1}$, satisfying the secant equation $\mathbf{J}^t \boldsymbol{\delta}^t = \boldsymbol{g}^t$, where

$$\boldsymbol{\delta}^t = \boldsymbol{\beta}^t - \boldsymbol{\beta}^{t-1} \quad \text{and} \quad \boldsymbol{g}^t = \nabla \widehat{Q}_h(\boldsymbol{\beta}^t) - \nabla \widehat{Q}_h(\boldsymbol{\beta}^{t-1}), \ \ t = 1, 2, \ldots. \tag{3.13}$$

To mitigate the computational cost of inverting a large matrix, the BB method chooses η so that $\eta \nabla \widehat{Q}_h(\boldsymbol{\beta}^t) = (\eta^{-1}\mathbf{I}_p)^{-1} \nabla \widehat{Q}_h(\boldsymbol{\beta}^t)$ "approximates" $(\mathbf{J}^t)^{-1} \nabla \widehat{Q}_h(\boldsymbol{\beta}^t)$. Since $\mathbf{J}^t$ satisfies $\mathbf{J}^t \boldsymbol{\delta}^t = \boldsymbol{g}^t$, it is more practical to choose η such that $(1/\eta)\boldsymbol{\delta}^t \approx \boldsymbol{g}^t$ or $\boldsymbol{\delta}^t \approx \eta \boldsymbol{g}^t$. Via least squares approximations, one may use $\eta_{1,t}^{-1} = \operatorname{argmin}_\alpha \|\alpha \boldsymbol{\delta}^t - \boldsymbol{g}^t\|_2^2$ or $\eta_{2,t} = \operatorname{argmin}_\eta \|\boldsymbol{\delta}^t - \eta \boldsymbol{g}^t\|_2^2$. The BB stepsizes are then defined as

$$\eta_{1,t} = \frac{\langle \boldsymbol{\delta}^t, \boldsymbol{\delta}^t \rangle}{\langle \boldsymbol{\delta}^t, \boldsymbol{g}^t \rangle} \quad \text{and} \quad \eta_{2,t} = \frac{\langle \boldsymbol{\delta}^t, \boldsymbol{g}^t \rangle}{\langle \boldsymbol{g}^t, \boldsymbol{g}^t \rangle}. \tag{3.14}$$

Consequently, the BB iteration takes the form

$$\boldsymbol{\beta}^{t+1} = \boldsymbol{\beta}^t - \eta_{\ell,t} \nabla \widehat{Q}_h(\boldsymbol{\beta}^t), \quad \ell = 1 \text{ or } 2. \tag{3.15}$$

Note that the BB step starts at iteration 1, while at iteration 0, we compute $\boldsymbol{\beta}^1$ using standard gradient descent with an initial estimate $\boldsymbol{\beta}^0$. The procedure is summarized in Algorithm 1. Based

Algorithm 1. Gradient descent with Barzilai-Borwein stepsize (GD-BB) for solving conquer.

Input: data vectors $\{(y_i, \boldsymbol{x}_i)\}_{i=1}^n$, $\tau \in (0,1)$, bandwidth $h \in (0,1)$, initialization $\boldsymbol{\beta}^0$, and gradient tolerance δ.

Compute $\boldsymbol{\beta}^1 \leftarrow \boldsymbol{\beta}^0 - \nabla \widehat{Q}_h(\boldsymbol{\beta}^0)$
for $t = 1, 2 \ldots$ **do**
  $\boldsymbol{\delta}^t \leftarrow \boldsymbol{\beta}^t - \boldsymbol{\beta}^{t-1}$, $\boldsymbol{g}^t \leftarrow \nabla \widehat{Q}_h(\boldsymbol{\beta}^t) - \nabla \widehat{Q}_h(\boldsymbol{\beta}^{t-1})$
  $\eta_{1,t} \leftarrow \langle \boldsymbol{\delta}^t, \boldsymbol{\delta}^t \rangle / \langle \boldsymbol{\delta}^t, \boldsymbol{g}^t \rangle$, $\eta_{2,t} \leftarrow \langle \boldsymbol{\delta}^t, \boldsymbol{g}^t \rangle / \langle \boldsymbol{g}^t, \boldsymbol{g}^t \rangle$
  $\eta_t \leftarrow \min\{\eta_{1,t}, \eta_{2,t}, 100\}$ if $\eta_{1,t} > 0$ and $\eta_t \leftarrow 1$ otherwise
  $\boldsymbol{\beta}^{t+1} \leftarrow \boldsymbol{\beta}^t - \eta_t \nabla \widehat{Q}_h(\boldsymbol{\beta}^t)$
end for when $\|\nabla \widehat{Q}_h(\boldsymbol{\beta}^t)\|_2 \leq \delta$

on extensive numerical studies, we find that at a fixed τ, the number of iterations is insensitive to varying (n, p) combinations. Moreover, as h increases, the number of iterations declines because the loss function is "more convex" for larger h. In Algorithm 1, the quantity $\delta > 0$ is a prespecified tolerance level, ensuring that the final iterate $\boldsymbol{\beta}^T$ satisfies $\|\nabla \widehat{Q}_h(\boldsymbol{\beta}^T)\|_2 \leq \delta$. Provided that $\delta \lesssim \sqrt{p/n}$, the statistical theory developed in Section 3.3 prevails. In our `R` package `conquer`, we set $\delta = 10^{-4}$ as the default value; this value can also be specified by the user.

As τ approaches 0 or 1, the Hessian matrix becomes more ill-conditioned. As a result, the stepsizes computed in GD-BB may sometimes fluctuate drastically, causing instability of the algorithm. Therefore, in practice, we set an upper bound for the stepsizes by taking $\eta_t = \min\{\eta_{1,t}, \eta_{2,t}, 100\}$, for $t = 1, 2, \ldots$. Another case of an ill-conditioned Hessian arises when we have covariates with very different scales. In this case, the stepsize should be different for each covariate, and a constant stepsize will be either too small or too large for one or more covariates, which leads to slow convergence. To address this issue, we scale the covariate inputs to have zero mean and unit variance before applying gradient descent.

3.2.2 Warm start via asymmetric Huber regression

A good initialization helps reduce the number of iterations for GD, and hence facilitates fast convergence. Recall from Remark 3.2.1 that with a uniform kernel, the smoothed check

function is proximal to a Huber loss [Huber, 1973]. Motivated by this subtle proximity, we propose using the asymmetric Huber M-estimator as an initial estimate, and then proceed by iteratively applying gradient descent with BB update step.

Let $H_{\tau,\gamma}(u) = |\tau - \mathbb{1}(u<0)| \cdot \{(u^2/2)\mathbb{1}(|u| \le \gamma) + \gamma(|u| - \gamma/2)\mathbb{1}(|u| > \gamma)\}$ be the asymmetric Huber loss parametrized by $\gamma > 0$. The asymmetric Huber M-estimator is then defined as

$$\widetilde{\boldsymbol{\beta}}_\gamma \in \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \widehat{\mathscr{L}_\gamma}(\boldsymbol{\beta}), \text{ where } \widehat{\mathscr{L}_\gamma}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} H_{\tau,\gamma}(y_i - \langle \boldsymbol{x}_i, \boldsymbol{\beta}\rangle). \tag{3.16}$$

The quantity γ is a shape parameter that controls the amount of robustness. The main reason for choosing a fixed (neither diminishing nor diverging) tuning parameter γ in Huber [1981] is to guarantee robustness towards arbitrary contamination in a neighborhood of the model. This is at the core of the robust statistics idiosyncrasy. In particular, Huber [1981] proposed $\gamma = 1.35\sigma$ to gain as much robustness as possible while retaining 95% asymptotic efficiency for normally distributed data, where $\sigma > 0$ is the standard deviation of the random noise. We estimate σ using the median absolute deviation of the residuals at each iteration, i.e., $\text{MAD}(\{r_i^t\}_{i=1}^n) = \text{median}(|r_i^t - \text{median}(r_i^t)|)$.

Noting that the asymmetric Huber loss is twice continuously differentiable, convex, and locally strongly convex, we use the GD-BB method described in the previous section to solve the optimization problem (3.16). Starting at iteration 0 with $\boldsymbol{\beta}^{0,0} = \mathbf{0}$, at iteration $t = 0, 1, 2, \ldots$, we compute

$$\boldsymbol{\beta}^{0,t+1} = \boldsymbol{\beta}^{0,t} - \eta_t \nabla \widehat{\mathscr{L}_\gamma}(\boldsymbol{\beta}^{0,t}) = \boldsymbol{\beta}^{0,t} + \frac{\eta_t}{n}\sum_{i=1}^{n} \psi_{\tau,\gamma}(y_i - \langle \boldsymbol{x}_i, \boldsymbol{\beta}^{0,t}\rangle)\boldsymbol{x}_i \tag{3.17}$$

with $\eta_t > 0$ automatically obtained by the BB method, where $\psi_{\tau,\gamma}(u) = |\tau - \mathbb{1}(u<0)| \cdot H'_{\tau,\gamma}(u) = |\tau - \mathbb{1}(u<0)| \cdot \min\{\max(-\gamma, u), \gamma\}$. The final iterate $\boldsymbol{\beta}^{0,T'}$ for some $T' > 1$ will be used as the initial value in Section 3.2.1. We summarize the details in Algorithm 2.

Algorithm 2. GD-BB method for solving (3.16).

Input: $\{(y_i, \boldsymbol{x}_i)\}_{i=1}^n$ and convergence criterion δ.

Initialize $\boldsymbol{\beta}^{0,0} = \mathbf{0}$

Compute $\gamma^0 = 1.35 \cdot \mathrm{MAD}(\{r_i^0\}_{i=1}^n)$, where $r_i^0 \leftarrow y_i - \langle \boldsymbol{x}_i, \boldsymbol{\beta}^{0,0} \rangle$, $i = 1, \ldots, n$, where $\mathrm{MAD}(\cdot)$ is the median absolute deviation

$\boldsymbol{\beta}^{0,1} \leftarrow \boldsymbol{\beta}^{0,0} - \nabla \widehat{\mathscr{L}_{\gamma^0}}(\boldsymbol{\beta}^{0,0})$

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

  $\gamma^t = 1.35 \cdot \mathrm{MAD}(\{r_i^t\}_{i=1}^n)$, where $r_i^t \leftarrow y_i - \langle \boldsymbol{x}_i, \boldsymbol{\beta}^{0,t} \rangle$, $i = 1, \ldots, n$

  $\boldsymbol{\delta}^t \leftarrow \boldsymbol{\beta}^{0,t} - \boldsymbol{\beta}^{0,t-1}$, $\boldsymbol{g}^t \leftarrow \nabla \widehat{\mathscr{L}_{\gamma^t}}(\boldsymbol{\beta}^{0,t}) - \nabla \widehat{\mathscr{L}_{\gamma^t}}(\boldsymbol{\beta}^{0,t-1})$

  $\eta_{1,t} \leftarrow \langle \boldsymbol{\delta}^t, \boldsymbol{\delta}^t \rangle / \langle \boldsymbol{\delta}^t, \boldsymbol{g}^t \rangle$, $\eta_{2,t} \leftarrow \langle \boldsymbol{\delta}^t, \boldsymbol{g}^t \rangle / \langle \boldsymbol{g}^t, \boldsymbol{g}^t \rangle$.

  $\eta_t \leftarrow \min\{\eta_{1,t}, \eta_{2,t}, 100\}$ if $\eta_{1,t} > 0$ and $\eta_t \leftarrow 1$ otherwise

  $\boldsymbol{\beta}^{0,t+1} \leftarrow \boldsymbol{\beta}^{0,t} - \eta_t \nabla \widehat{\mathscr{L}_{\gamma^t}}(\boldsymbol{\beta}^{0,t})$

end for when $\|\nabla \widehat{\mathscr{L}_{\gamma^t}}(\boldsymbol{\beta}^{0,t})\|_2 \leq \delta$

3.3 Statistical Analysis

Under the linear quantile regression model, we write, for convenience, the generic data vector $(y, \boldsymbol{x})$ in a linear model form: given a quantile level $\tau \in (0,1)$ of interest,

$$y = \langle \boldsymbol{x}, \boldsymbol{\beta}^*(\tau) \rangle + \varepsilon(\tau), \tag{3.18}$$

where the random variable $\varepsilon(\tau)$ satisfies $\mathbb{P}\{\varepsilon(\tau) \leq 0 | \boldsymbol{x}\} = \tau$. Let $f_{\varepsilon|\boldsymbol{x}}(\cdot)$ be the conditional density function of the regression error $\varepsilon = \varepsilon(\tau)$ given $\boldsymbol{x} = (x_1, \ldots, x_p)^{\mathrm{T}}$ ($p \geq 2$). We first derive upper bounds for the smoothing bias under mild regularity conditions on the conditional density $f_{\varepsilon|\boldsymbol{x}}$ and the kernel function. For any vector $\boldsymbol{u} \in \mathbb{R}^p$, we write $\boldsymbol{u}_- \in \mathbb{R}^{p-1}$ as the sub-vector of $\boldsymbol{u}$ with its first component removed. Recall that $x_1 \equiv 1$, and $\boldsymbol{x}_- = (x_2, \ldots, x_p)^{\mathrm{T}} \in \mathbb{R}^{p-1}$ is assumed to be random. Without loss of generality, we assume $\boldsymbol{\mu}_- := \mathbb{E}(\boldsymbol{x}_-) = \mathbf{0}$ throughout this section; otherwise, set $\widetilde{\boldsymbol{x}} = (1, (\boldsymbol{x}_- - \boldsymbol{\mu}_-)^{\mathrm{T}})^{\mathrm{T}}$, so that model (3.18) can be written as $y = \langle \widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{\beta}}^* \rangle + \varepsilon$, where $\widetilde{\boldsymbol{\beta}}^* = (\widetilde{\beta}_1^*, \beta_2^*, \ldots, \beta_p^*)^{\mathrm{T}}$ with $\widetilde{\beta}_1^* = \beta_1^* + \langle \boldsymbol{\mu}_-, \boldsymbol{\beta}_-^* \rangle$. The analysis then applies to $\{(y_i, \widetilde{\boldsymbol{x}}_i)\}_{i=1}^n$, and the probabilistic bounds for $\widetilde{\boldsymbol{\beta}}^*$ naturally lead to those for $\boldsymbol{\beta}^*$.

3.3.1 Smoothing bias

Condition 3.3.1 (Kernel function)**.** *Let $K(\cdot)$ be a symmetric and non-negative function that integrates to one, that is, $K(u) = K(-u)$, $K(u) \geq 0$ for all $u \in \mathbb{R}$ and $\int_{-\infty}^{\infty} K(u)\,\mathrm{d}u = 1$. Moreover, $K(\cdot)$ is bounded with $\kappa_u := \sup_{u\in\mathbb{R}} K(u) < \infty$.*

We will use the notation $\kappa_k = \int_{-\infty}^{\infty} |u|^k K(u)\,\mathrm{d}u$ for $k \geq 1$. Furthermore, we define the population smoothed loss function $Q_h(\boldsymbol{\beta}) = \mathbb{E}\{\widehat{Q}_h(\boldsymbol{\beta})\}$, $\boldsymbol{\beta} \in \mathbb{R}^p$ and the pseudo parameter

$$\boldsymbol{\beta}_h^*(\tau) \in \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} Q_h(\boldsymbol{\beta}), \tag{3.19}$$

which is the population minimizer under the smoothed quantile loss. For simplicity, we write $\boldsymbol{\beta}^* = \boldsymbol{\beta}^*(\tau)$ and $\boldsymbol{\beta}_h^* = \boldsymbol{\beta}_h^*(\tau)$ hereinafter. In general, $\boldsymbol{\beta}_h^*$ differs from $\boldsymbol{\beta}^*$, and we refer to $\|\boldsymbol{\beta}_h^* - \boldsymbol{\beta}^*\|_2$ as the approximation error or smoothing bias.

Condition 3.3.2 (Conditional density)**.** *There exists $\underline{f} > 0$ such that $f_{\varepsilon|\boldsymbol{x}}(0) \geq \underline{f}$ almost surely (for all $\boldsymbol{x}$). Moreover, there exists a constant $l_0 > 0$ such that $|f_{\varepsilon|\boldsymbol{x}}(u) - f_{\varepsilon|\boldsymbol{x}}(v)| \leq l_0|u-v|$ for all $u, v \in \mathbb{R}$ almost surely (over $\boldsymbol{x}$).*

Condition 3.3.3 (Random design: moments)**.** *The (random) vector $\boldsymbol{x} \in \mathbb{R}^p$ of covariates satisfies $m_3 := \sup_{\boldsymbol{u}\in\mathbb{S}^{p-1}} \mathbb{E}(|\langle \boldsymbol{u}, \boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}\rangle|^3) < \infty$, where $\boldsymbol{\Sigma} = \mathbb{E}(\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}})$ is positive definite.*

Condition 3.3.3 requires that all the one-dimensional marginals of $\boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}$ have bounded third absolute moments. When $\boldsymbol{x}_-$ follows a multivariate normal distribution, Condition 3.3.3 holds trivially. The following result characterizes the smoothing bias from a non-asymptotic viewpoint.

Proposition 3.3.1. *Assume Conditions 3.3.1–3.3.3 hold, and let the bandwidth satisfy $0 < h < \frac{1}{l_0\{\kappa_1 + (m_3\kappa_2)^{1/2}\}}\underline{f}$. Then, $\boldsymbol{\beta}_h^*$ is the unique minimizer of $\boldsymbol{\beta} \mapsto Q_h(\boldsymbol{\beta})$ and satisfies*

$$\delta_h := \|\boldsymbol{\beta}_h^* - \boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma}} < \frac{l_0\kappa_2 h^2}{\underline{f} - l_0\kappa_1 h}. \tag{3.20}$$

In addition, assume $f_{\varepsilon|\boldsymbol{x}}(\cdot)$ is continuously differentiable and satisfies almost surely (over $\boldsymbol{x}$) that $|f'_{\varepsilon|\boldsymbol{x}}(u) - f'_{\varepsilon|\boldsymbol{x}}(0)| \leq l_1|u|$ for some constant $l_1 > 0$. Then

$$\left\| \boldsymbol{\Sigma}^{-1/2}\mathbf{J}(\boldsymbol{\beta}_h^* - \boldsymbol{\beta}^*) + \frac{1}{2}\kappa_2 h^2 \cdot \boldsymbol{\Sigma}^{-1/2}\mathbb{E}\big\{f'_{\varepsilon|\boldsymbol{x}}(0)\boldsymbol{x}\big\} \right\|_2 \leq \frac{1}{6}l_1\kappa_3 h^3 + \frac{1}{2}l_0 m_3 \delta_h^2 + l_0\kappa_1 h\delta_h, \quad (3.21)$$

where $\mathbf{J} = \mathbb{E}\{f_{\varepsilon|\boldsymbol{x}}(0)\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\}$.

To better understand the bounds (3.20) and (3.21), note that $\|\boldsymbol{\beta}_h^* - \boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma}}^2 = \mathbb{E}\langle \boldsymbol{x}, \boldsymbol{\beta}_h^* - \boldsymbol{\beta}^*\rangle^2$ is the average prediction smoothing error. Interestingly, the upper bound on the right-hand side is dimension-free given h as long as the uniform third moment m_3 in Condition 3.3.3 is dimension-free. Another interesting implication is that, when both $f_{\varepsilon|\boldsymbol{x}}(0)$ and $f'_{\varepsilon|\boldsymbol{x}}(0)$ are independent of $\boldsymbol{x}$, i.e., $f_{\varepsilon|\boldsymbol{x}}(0) = f_\varepsilon(0)$ and $f'_{\varepsilon|\boldsymbol{x}}(0) = f'_\varepsilon(0)$, the leading term in the bias simplifies to

$$\frac{1}{2}\kappa_2 h^2 \cdot \mathbf{J}^{-1}\mathbb{E}\big\{f'_{\varepsilon|\boldsymbol{x}}(0)\boldsymbol{x}\big\} = \frac{f'_\varepsilon(0)}{2f_\varepsilon(0)}\kappa_2 h^2 \cdot \boldsymbol{\Sigma}^{-1}\mathbb{E}(\boldsymbol{x}) = \frac{f'_\varepsilon(0)}{2f_\varepsilon(0)}\kappa_2 h^2 \cdot \begin{bmatrix} 1 \\ \mathbf{0}_{p-1} \end{bmatrix}.$$

In other words, the smoothing bias is concentrated primarily on the intercept. In the asymptotic setting where p is fixed, and $h = o(1)$ as $n \to \infty$, we refer to Theorem 1 in Fernandes, Guerre and Horta [2021] for the expression of asymptotic bias.

3.3.2 Finite sample theory

In this section, we provide two non-asymptotic results, the concentration inequality and the Bahadur-Kiefer representation, for the conquer estimator under random design.

Condition 3.3.4 (Random design: sub-exponential case)**.** *The predictor $\boldsymbol{x} = (x_1, \ldots, x_p)^{\mathrm{T}} \in \mathbb{R}^p$ is sub-exponential with $x_1 \equiv 1$ and $\mathbb{E}(x_j) = 0$ for $j = 2, \ldots, p$. That is, there exists $\upsilon_0 > 0$ such that $\mathbb{P}\{|\langle \boldsymbol{u}, \boldsymbol{w}\rangle| \geq \upsilon_0 t\} \leq e^{-t}$ for all $\boldsymbol{u} \in \mathbb{S}^{p-1}$ and $t \geq 0$, where $\boldsymbol{w} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}$ with $\boldsymbol{\Sigma} = \mathbb{E}(\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}})$ being positive definite.*

Condition 3.3.4 asserts that the distribution of the covariates is sub-exponential, which

encompasses the bounded case considered by Fernandes, Guerre and Horta [2021]. For the standardized predictor $\boldsymbol{w} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}$, we define the uniform moment parameters (including m_3 that first occurred in Condition (3.3.3))

$$m_k = \sup_{\boldsymbol{u}\in\mathbb{S}^{p-1}} \mathbb{E}|\langle \boldsymbol{u}, \boldsymbol{w}\rangle|^k, \ \ k = 1, 2, \ldots, \tag{3.22}$$

with $m_2 = 1$. In particular, m_4 can be viewed as the uniform kurtosis parameter. Under Condition 3.3.4, a straightforward calculation shows that $m_k \le \upsilon_0^k k!$, valid for all $k \ge 1$.

Theorem 3.3.1. *Assume Conditions 3.3.1, 3.3.2 and 3.3.4 hold. For any $t > 0$, the smoothed quantile regression estimator $\widehat{\boldsymbol{\beta}}_h$ with $\underline{f}^{-1} m_3^{1/2} \upsilon_0 \sqrt{(p+t)/n} \lesssim h \lesssim \underline{f} m_3^{-1/2}$ satisfies the bound*

$$\|\widehat{\boldsymbol{\beta}}_h - \boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma}} \lesssim \frac{1}{\underline{f}} \left\{ \upsilon_0 \sqrt{\frac{\log_2(1/h) + p + t}{n}} + l_0 \kappa_2 h^2 \right\}, \tag{3.23}$$

with probability at least $1 - 2e^{-t}$, where $\log_2(x) := \log\log(x \vee 1)$.

With high probability, the estimation error in (3.23) is upper bounded by two terms, $\underline{f}^{-1} l_0 \kappa_2 h^2$ and $\underline{f}^{-1} \upsilon_0 \sqrt{(p+t)/n}$, which can be interpreted as the bias and statistical rate of convergence, respectively. The parameter $t \ge 0$ controls the confidence level through $1 - 2e^{-t}$. The additional factor $\log_2(1/h)$ in the upper bound is a consequence of the peeling argument, which can be removed via a more refined analysis yet under slightly stronger technical conditions; see Section B.2.3 in the supplement for details. Adjusting the proof by changing high probability bounds to $\mathscr{O}_{\mathbb{P}}$ statements, it can be shown that $\|\widehat{\boldsymbol{\beta}}_h - \boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma}} = \mathscr{O}_{\mathbb{P}}(\sqrt{p/n})$ under the condition $h = \mathscr{O}((p/n)^{1/4})$ and $\sqrt{p/n} = \mathscr{O}(h)$. Next we explain the bandwidth constraint $\sqrt{p/n} \lesssim h \lesssim 1$ required in Theorem 3.3.1 and all the other results below. On the one side, the smoothing parameter should be sufficiently small, typically $h = h_n \to 0$, so that the smoothing bias is negligible and does not change the target parameter to be estimated. On the other side, the bandwidth cannot be too small in the sense that we need $h \gtrsim \sqrt{p/n}$. Intuitively, this is because the main motivation for smoothed QR is to seek a tradeoff between statistical rate of convergence and

computational precision (unless the data is noiseless). The standard QR estimator $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}(\tau)$ has a convergence rate $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 = \mathscr{O}_{\mathbb{P}}(\sqrt{p/n})$ under the growth condition $p \sup_{\boldsymbol{x} \in \mathscr{X}} \|\boldsymbol{x}\|_2^2 \cdot (\log n)^2 = o(n)$; see Theorem 1 in Belloni et al. [2019]. Here $\mathscr{X} \subseteq \mathbb{R}^p$ is the support of the covariate vector $\boldsymbol{x} \in \mathbb{R}^p$. Therefore, smoothing will become redundant if the bandwidth is set at a level below the best possible statistical convergence radius.

Our results provide non-asymptotic bounds via high probability statements, which complement the classical Big OP ($\mathscr{O}_{\mathbb{P}}$) and little op ($o_{\mathbb{P}}$) statements frequently used in statistics and econometrics. Probabilistic bounds of this kind can also be extended to analyze high-dimensional models [Belloni and Chernozhukov, 2011, Wang, Wu and Li, 2012] or nonparametric methods [Belloni et al., 2019].

Next, we establish a Bahadur representation for the conquer estimator, which lays the theoretical foundation for the ensuing statistical inference. To this end, we impose a slightly more stringent sub-Gaussian condition on the covariates, where the sub-Gaussian parameter is a dimension-free constant.

Condition 3.3.5 (Random design: sub-Gaussian case)**.** *The predictor $\boldsymbol{x} = (x_1, \ldots, x_p)^{\mathrm{T}} \in \mathbb{R}^p$ is sub-Gaussian with $x_1 \equiv 1$ and $\mathbb{E}(x_j) = 0$ for $j = 2, \ldots, p$. That is, there exists $\upsilon_1 > 0$ such that $\mathbb{P}\{|\langle \boldsymbol{u}, \boldsymbol{w} \rangle| \geq \upsilon_1 t\} \leq 2e^{-t^2/2}$ for all $\boldsymbol{u} \in \mathbb{S}^{p-1}$ and $t \geq 0$, where $\boldsymbol{w} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}$.*

Theorem 3.3.2. *In addition to Conditions 3.3.1, 3.3.2 and 3.3.5, assume $\sup_{u \in \mathbb{R}} f_{\varepsilon|\boldsymbol{x}}(u) \leq \bar{f}$ almost surely. Let $t > 0$, and suppose n and h satisfy $\underline{f}^{-1} m_3^{1/2} \upsilon_1 \sqrt{(p+t)/n} \lesssim h \lesssim \underline{f} m_3^{-1/2}$. Then, with probability at least $1 - 3e^{-t}$,*

$$\left\| \boldsymbol{\Sigma}^{-1/2} \mathbf{J}_h (\widehat{\boldsymbol{\beta}}_h - \boldsymbol{\beta}^*) - \frac{1}{n} \sum_{i=1}^{n} \big\{ \tau - \mathscr{K}_h(-\varepsilon_i) \big\} \boldsymbol{\Sigma}^{-1/2} \boldsymbol{x}_i \right\|_2 \lesssim \frac{p+t}{nh^{1/2}} + h^{3/2} \sqrt{\frac{p+t}{n}} + h^4, \quad (3.24)$$

where $\mathbf{J}_h = \nabla^2 Q_h(\boldsymbol{\beta}^*) = \mathbb{E}\big\{K_h(\varepsilon)\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\big\}$, $\mathscr{K}_h(u) = \int_{-\infty}^{u/h} K(v)\,\mathrm{d}v$. When $\mathbf{J}_h$ on the left-hand side of (3.24) *is replaced by $\mathbf{J} = \mathbb{E}\{f_{\varepsilon|\boldsymbol{x}}(0)\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\}$, the upper bound is of order $(p+t)/(nh^{1/2}) + h\sqrt{(p+t)/n} + h^3$.*

With growing dimensions (many regressors), Theorem 3.3.2 is directly comparable to and complements Theorem 2 in Belloni et al. [2019], although the latter concerns the linear approximation of the quantile regression process. To see the connection, we write $n^{1/2}(\widehat{\boldsymbol{\beta}}_h - \boldsymbol{\beta}^*) = \mathbf{J}_h^{-1}\boldsymbol{U}_h + \boldsymbol{r}_h$, where $\boldsymbol{U}_h = n^{-1/2}\sum_{i=1}^n (1-\mathbb{E})\{\tau - \mathscr{K}_h(-\varepsilon_i)\}\boldsymbol{x}_i$ is a zero-mean random vector, and the remainder $\boldsymbol{r}_h$ is such that $\|\boldsymbol{r}_h\|_2 \lesssim (p+t)/(nh)^{1/2} + n^{1/2}h^2$ with high probability. Minimizing the right-hand side over h in terms of order leads to a convergence rate $p^{4/5}/n^{3/10}$. For standard QR with fixed design, Theorem 2 in Belloni et al. [2019] implies $n^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) = \mathbf{J}^{-1}\boldsymbol{U} + \boldsymbol{r}$, where $\boldsymbol{U} = n^{-1/2}\sum_{i=1}^n \{\tau - \mathbb{1}(\varepsilon_i \le 0)\}\boldsymbol{x}_i$, and $\|\boldsymbol{r}\|_2 = \mathscr{O}_{\mathbb{P}}(p^{3/4}\zeta_p(\log n)^{1/2}n^{-1/4})$, where $\zeta_p = \sup_{\boldsymbol{x}\in\mathscr{X}}\|\boldsymbol{x}\|_2$. From an asymptotic perspective, the QR estimator has the advantage of being (conditionally) pivotal asymptotically. However, possibly due to the non-smoothness of the check function, the linear approximation error has a slower rate of convergence $(p^5/n)^{1/4}$ even for bounded design, i.e., $\zeta_p \le Bp^{1/2}$ for some constant $B > 0$. For the conquer estimator, although the linear term $\boldsymbol{U}_h$ is not pivotal, we will show that Rademacher multiplier bootstrap provides accurate approximations both theoretically and numerically.

The Bahadur representation can be used to establish the limiting distribution of the estimator or its functionals. Here we consider a fundamental statistical inference problem for testing the linear hypothesis $H_0 : \langle \boldsymbol{a}, \boldsymbol{\beta}^* \rangle = 0$, where $\boldsymbol{a} \in \mathbb{R}^p$ is a deterministic vector that defines a linear functional of interest. It is then natural to consider a test statistic that depends on $n^{1/2}\langle \boldsymbol{a}, \widehat{\boldsymbol{\beta}}_h \rangle$. Based on the non-asymptotic result in Theorem 3.3.2, we establish a Berry-Esseen bound for the linear projection of the conquer estimator.

Theorem 3.3.3. *Assume that the conditions in Theorem 3.3.2 hold, and* $\sqrt{(p+\log n)/n} \lesssim h \lesssim 1$. *Then,*

$$\Delta_{n,p}(h) := \sup_{x\in\mathbb{R},\,\boldsymbol{a}\in\mathbb{R}^p} \left| \mathbb{P}\big(n^{1/2}\sigma_h^{-1}\langle \boldsymbol{a}, \widehat{\boldsymbol{\beta}}_h - \boldsymbol{\beta}^* \rangle \le x\big) - \Phi(x) \right| \lesssim \frac{p+\log n}{(nh)^{1/2}} + n^{1/2}h^2, \qquad (3.25)$$

where $\sigma_h^2 = \sigma_h^2(\boldsymbol{a}) = \boldsymbol{a}^{\mathrm{T}}\mathbf{J}_h^{-1}\mathbb{E}\big[\{\mathscr{K}_h(-\varepsilon) - \tau\}^2\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\big]\mathbf{J}_h^{-1}\boldsymbol{a}$, *where* $\Phi(\cdot)$ *denotes the standard nor-*

mal distribution function. Moreover,

$$\sup_{\boldsymbol{a}\in\mathbb{R}^p}\left|\frac{\sigma_h^2(\boldsymbol{a})}{\boldsymbol{a}^{\mathrm{T}}\mathbf{J}_h^{-1}\boldsymbol{\Sigma}\mathbf{J}_h^{-1}\boldsymbol{a}}-\tau(1-\tau)\right|=O(h)\ \ \textit{as } h\to 0.$$

If, in addition, that $f_{\varepsilon|\boldsymbol{x}}(\cdot)$ is twice continuously differentiable and satisfies $|f''_{\varepsilon|\boldsymbol{x}}(u)-f''_{\varepsilon|\boldsymbol{x}}(v)|\leq l_2(\boldsymbol{x})|u-v|$ for all $u,v\in\mathbb{R}$ and $\boldsymbol{x}\in\mathbb{R}^p$, and $l_2:\mathbb{R}^p\to\mathbb{R}^+$ is such that $\mathbb{E}\{l_2^2(\boldsymbol{x})\}\leq C$ for some $C>0$. Then,

$$\begin{aligned}\sup_{x\in\mathbb{R},\boldsymbol{a}\in\mathbb{R}^p}\left|\mathbb{P}\big(n^{1/2}\sigma_h^{-1}\langle\boldsymbol{a},\widehat{\boldsymbol{\beta}}_h-\boldsymbol{\beta}^*+0.5\kappa_2 h^2\mathbf{J}_h^{-1}\mathbb{E}\{f'_{\varepsilon|\boldsymbol{x}}(0)\boldsymbol{x}\}\rangle\leq x\big)-\Phi(x)\right|&\\ \lesssim\frac{p+\log n}{(nh)^{1/2}}+(p+\log n)^{1/2}h^{3/2}+n^{1/2}h^4.&\end{aligned}\tag{3.26}$$

Theorem 2.1.3 shows that for certain choice of bandwidth $h=h_n\to 0$, all the linear functionals of $\widehat{\boldsymbol{\beta}}_h$, after properly standardization, are asymptotically normal as $n,p\to\infty$ subject to some conditions. For example, if h satisfies $h=o(n^{-1/4})$, then the smoothing bias does not affect the asymptotic distribution. The Berry-Esseen bound (3.25) immediately yields a large-p asymptotic result. Taking $h=h_n=\{(p+\log n)/n\}^{2/5}$ therein, the Gaussian approximation error $\Delta_{n,p}(h)$ is of order $(p+\log n)^{4/5}n^{-3/10}$. Consequently, $n^{1/2}\langle\boldsymbol{a},\widehat{\boldsymbol{\beta}}_h-\boldsymbol{\beta}^*\rangle$, for any given (deterministic) vector $\boldsymbol{a}\in\mathbb{R}^p$, is asymptotically normally distributed as long as $p^{8/3}/n\to 0$, which improves the best known growth condition on p for quantile regression [Welsh, 1989].

Remark 3.3.1 (Large-p asymptotics)**.** *This is a follow-up discussion of Remark 2.1.1. For smooth robust regression estimators, asymptotic normality can be proven under less restrictive conditions on p. Huber [1973] showed that if the loss is twice differentiable, the asymptotic normality for $\langle\boldsymbol{a},\widehat{\boldsymbol{\beta}}\rangle$, where $\boldsymbol{a}\in\mathbb{R}^p$, holds if $p^3/n\to 0$ as n increases. Portnoy [1985] and Mammen [1989] weakened this condition to $(p\log n)^{3/2}/n\to 0$ and $p^{3/2}\log(n)/n\to 0$, respectively, when the loss function is four times differentiable. For Huber loss that has a Lipschitz continuous derivative, He and Shao [2000] obtained the scaling $p^2\log p=o(n)$ that ensures the asymptotic normality of arbitrary linear combinations of $\widehat{\boldsymbol{\beta}}$. Table 1 summarizes our discussion here and shows that*

the smoothing for conquer helps ensure asymptotic normality of the estimator under weaker conditions on p than what we need for the usual quantile regression estimator.

Table 3.1. Summary of scaling conditions required for normal approximation under various loss functions.

Loss function	Design	Scaling condition
Huber loss [Huber, 1973]	Fixed design	$p^3 = o(n)$
Four times differentiable loss [Portnoy, 1985]	Fixed design (with symmetric error)	$(p\log n)^{3/2} = o(n)$
Four times differentiable loss [Mammen, 1989]	Fixed design	$p^{3/2}\log n = o(n)$
Huber loss [He and Shao, 2000]	Fixed design	$p^2 \log p = o(n)$
Huber loss [Chen and Zhou, 2020]	Sub-Gaussian	$p^2 = o(n)$
Quantile loss [Welsh, 1989, He and Shao, 2000]	Fixed design	$p^3(\log n)^2 = o(n)$
Quantile loss [Pan and Zhou, 2021]	Sub-Gaussian	$p^3(\log n)^2 = o(n)$
Convolution smoothed quantile loss [He et al., 2022]	Sub-Gaussian	$p^{8/3} = o(n)$

Remark 3.3.2. *In this work, we show that the accuracy of conquer-based inference via the Bahadur representation (and normal approximations) has an error of rate faster than $n^{-1/4}$ yet slower than $n^{-1/2}$; see Theorems 3.3.2 and 3.3.3. For standard regression quantiles, Portnoy [2012] proposed an alternative expansion for the quantile process using the "Hungarian" construction of Komlós, Major and Tusnády. This stochastic approximation yields an error of order $n^{-1/2}$ (up to a factor of* $\log n$*), and hence provides a theoretical justification for accurate approximations for inference in regression quantile models.*

3.3.3 Theoretical guarantees for inference

We next investigate the statistical properties of the Rademacher multiplier bootstrap (RMB) defined in (3.9). As before, we consider array (non)asymptotics, and the obtained bootstrap approximation errors depend explicitly on (n, p) and h.

Theorem 3.3.4. *Assume Conditions 3.3.1, 3.3.2 and 3.3.5 hold. For any given $t \geq 0$, let the sample size and bandwidth satisfy $\underline{f}^{-1} m_3^{1/2} \upsilon_1 \sqrt{(p+t)/n} \lesssim h \lesssim \underline{f} m_3^{-1/2}$. Then, there exists some "good" event $\mathscr{E}(t)$ with $\mathbb{P}\{\mathscr{E}(t)\} \geq 1 - 3e^{-t}$ such that, with $\mathbb{P}^*$-probability at least $1 - 2e^{-t}$*

conditioned on $\mathscr{E}(t)$,

$$\|\widehat{\boldsymbol{\beta}}_h^{\flat} - \boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma}} \lesssim \frac{1}{\underline{f}} \left\{ \upsilon_1 \sqrt{\frac{\log_2(1/h) + p + t}{n}} + l_0 \kappa_2 h^2 \right\}. \tag{3.27}$$

Analogously to Theorem 3.3.2, we further provide a Bahadur representation result for the bootstrap estimator $\widehat{\boldsymbol{\beta}}_h^{\flat}$, which paves the way for validating the conquer-RMB method.

Theorem 3.3.5. *In addition to Conditions 3.3.1, 3.3.2 and 3.3.5, assume* $\sup_{u\in\mathbb{R}} f_{\varepsilon|\boldsymbol{x}}(u) \le \bar{f}$ *almost surely (in* $\boldsymbol{x}$*) and* $K(\cdot)$ *is* l_K*-Lipschitz continuous. Suppose the sample size satisfies* $n \gtrsim q := p + \log n$*, and set the bandwidth as* $h \asymp (q/n)^{2/5}$*. Then, there exists a sequence of events* $\{\mathscr{F}_n\}$ *with* $\mathbb{P}(\mathscr{F}_n) \ge 1 - 6n^{-1}$ *such that, with* $\mathbb{P}^*$*-probability at least* $1 - 3n^{-1}$ *conditioned on* $\mathscr{F}_n$*,*

$$\begin{aligned} &\left\| \boldsymbol{\Sigma}^{-1/2} \mathbf{J}_h (\widehat{\boldsymbol{\beta}}_h^{\flat} - \widehat{\boldsymbol{\beta}}_h) - \frac{1}{n} \sum_{i=1}^n e_i \{\tau - \mathscr{K}_h(-\varepsilon_i)\} \boldsymbol{\Sigma}^{-1/2} \boldsymbol{x}_i \right\|_2 \\ &\qquad \lesssim \left(\frac{q}{n}\right)^{4/5} \bigvee \left(\frac{q}{n}\right)^{3/5} \left(\frac{p \log n}{n}\right)^{1/4} \bigvee \left(\frac{q}{n}\right)^{3/5} \frac{p \log n}{n^{1/2}}. \end{aligned} \tag{3.28}$$

As suggested by Theorem 2.1.3 and the discussion below, if we set the order of the bandwidth h as $\{(p+\log n)/n\}^{2/5}$, the normal approximation to the conquer estimator is asymptotically accurate provided that $p^{8/3} = o(n)$ as $n \to \infty$. For the same h, the right-hand side of (3.28) is of order $o(n^{-1/2})$ provided that $p^{8/3}(\log n)^{5/3} = o(n)$. Putting these two parts together, we have the following asymptotic bootstrap approximation result.

Corollary 3.3.1. *Assume the same conditions of Theorem 3.3.5, and let the bandwidth be of order* $h \asymp \{(p+\log n)/n\}^{2/5}$*. If the dimension* $p = p_n$ *is subject to* $p(\log n)^{5/8} = o(n^{3/8})$*, then for any deterministic vector* $\boldsymbol{a} \in \mathbb{R}^p$*,*

$$\sup_{x\in\mathbb{R}} \left| \mathbb{P}\big(n^{1/2} \langle \boldsymbol{a}, \widehat{\boldsymbol{\beta}}_h - \boldsymbol{\beta}^* \rangle \le x\big) - \mathbb{P}^*\big(n^{1/2} \langle \boldsymbol{a}, \widehat{\boldsymbol{\beta}}_h^{\flat} - \widehat{\boldsymbol{\beta}}_h \rangle \le x\big) \right| \xrightarrow{\mathbb{P}} 0 \ \text{ as } n \to \infty. \tag{3.29}$$

The proof of (3.29) follows the same argument as that in the proof of Theorem 3.3.3, and therefore is omitted. The additional logarithmic factor in the scaling may be an artifact of the proof technique.

Remark 3.3.3. *(Multiplier bootstrap with more general weighting schemes) By examining the proof of Theorem 3.3.5, we see that the assumption* $\mathbb{E}(e_i^2) = 1$ *is not necessarily required for the bound* (3.28) *on bootstrap Bahadur linearization error. To retain the convexity of the bootstrap loss* $\widehat{Q}_h^\flat(\boldsymbol{\beta}) = (1/n)\sum_{i=1}^n (1+e_i)\ell_h(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})$*, we restrict our attention to non-negative multipliers* $1+e_i \geq 0$*. More generally, assume that* $e_1, \ldots, e_n$ *are i.i.d. satisfying*

$$\mathbb{E}(e_i) = 0, \quad e_i \geq -1 \ \text{ and } \ \log \mathbb{E} e^{\lambda e_i} \leq \lambda^2 \nu / 2 \ \text{ for all } \lambda \geq 0 \text{ and some } \nu > 0. \tag{3.30}$$

This means that e_i *has sub-Gaussian right tails. Typical examples satisfying* (3.30) *include: (i) uniform distribution on* $[-1,1]$*, (ii) symmetric triangular distribution on* $[-1,1]$*, (iii) shifted folded normal distribution* $(\pi/2)^{1/2}|g| - 1$ *where* $g \sim \mathcal{N}(0,1)$*. The proof of the bound* (3.28) *under such a general scheme requires more involved argument; see, for example, the proof of Theorem 2.3 in Chen and Zhou [2020] (the unit variance assumption therein can also be relaxed). When* $\kappa^2 := \mathbb{E}(e_i^2) \neq 1$*, although the bootstrap approximation result* (3.29) *will no longer hold, by a simple variance adjustment it can be shown that*

$$\sup_{x \in \mathbb{R}} \left| \mathbb{P}\big(n^{1/2}\langle \boldsymbol{a}, \widehat{\boldsymbol{\beta}}_h - \boldsymbol{\beta}^* \rangle \leq x\big) - \mathbb{P}^*\left\{ (n/\kappa)^{1/2} \langle \boldsymbol{a}, \widehat{\boldsymbol{\beta}}_h^\flat - \widehat{\boldsymbol{\beta}}_h \rangle \leq x \right\} \right| \xrightarrow{\mathbb{P}} 0 \ \text{ as } n \to \infty.$$

The pivotal bootstrap confidence intervals can thus be constructed by slightly adapting the method described in Section 3.1.3.

The numerical performance of the Rademacher multiplier bootstrap inference for conquer will be examined in Section 3.4.2. The main advantage of the multiplier bootstrap method is that it does not require estimating the variance-covariance matrices in (3.10), which can be quite unstable and thus causes outliers when τ is close to 0 or 1.

The construction of normal-based confidence intervals is based on the estimated variances $\widehat{\sigma}_h^2(\boldsymbol{a}) = \boldsymbol{a}^{\mathrm{T}}\widehat{\mathbf{J}}_h^{-1}\widehat{\mathbf{V}}_h\widehat{\mathbf{J}}_h^{-1}\boldsymbol{a}$ for $\boldsymbol{a} \in \mathbb{R}^p$, where $\widehat{\mathbf{J}}_h$ and $\widehat{\mathbf{V}}_h$ are given in (3.10). In view of Theorem 3.3.3, the validity of normal calibration relies on the consistency of $\widehat{\mathbf{J}}_h$ and $\widehat{\mathbf{V}}_h$. In the following, we provide the consistency of $\widehat{\mathbf{J}}_h$ and $\widehat{\mathbf{V}}_h$ under the operator norm, again in the regime "$p/n \to 0$ as $p, n \to \infty$".

Note that both $\widehat{\mathbf{J}}_h$ and $\widehat{\mathbf{V}}_h$ depend on the conquer estimator, whose rate of convergence is already established in Theorem 3.3.1. For $\boldsymbol{\delta} \in \mathbb{R}^p$, define matrix-valued functions

$$\widehat{\mathbf{J}}_h(\boldsymbol{\delta}) = \frac{1}{n}\sum_{i=1}^n K_h(\varepsilon_i - \langle \boldsymbol{x}_i, \boldsymbol{\delta}\rangle)\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}} \quad \text{and} \quad \widehat{\mathbf{V}}_h(\boldsymbol{\delta}) = \frac{1}{n}\sum_{i=1}^n \{\mathscr{K}_h(\langle \boldsymbol{x}_i, \boldsymbol{\delta}\rangle - \varepsilon) - \tau\}^2\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}, \quad (3.31)$$

so that $\widehat{\mathbf{J}}_h = \widehat{\mathbf{J}}_h(\widehat{\boldsymbol{\delta}})$ and $\widehat{\mathbf{V}}_h = \widehat{\mathbf{V}}_h(\widehat{\boldsymbol{\delta}})$ with $\widehat{\boldsymbol{\delta}} = \widehat{\boldsymbol{\beta}}_h - \boldsymbol{\beta}^*$. Conditioned on the event $\{\|\widehat{\boldsymbol{\delta}}\|_{\boldsymbol{\Sigma}} \le r\}$ for some prespecified $r > 0$ which determines the convergence rate of $\widehat{\boldsymbol{\beta}}_h$, we have

$$\|\widehat{\mathbf{J}}_h - \mathbf{J}_h\|_2 \le \sup_{\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}} \le r} \|\widehat{\mathbf{J}}_h(\boldsymbol{\delta}) - \mathbf{J}_h\|_2 \quad \text{and} \quad \|\widehat{\mathbf{V}}_h - \mathbf{V}_h\|_2 \le \sup_{\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}} \le r} \|\widehat{\mathbf{V}}_h(\boldsymbol{\delta}) - \mathbf{V}_h\|_2,$$

where $\mathbf{V}_h := \mathbb{E}\big[\{\mathscr{K}_h(-\varepsilon) - \tau\}^2\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\big]$. The problem is thus reduced to controlling the above suprema over a local neighborhood.

Proposition 3.3.2. *In addition to the conditions in Theorem 3.3.2, assume that the kernel $K(\cdot)$ is l_K-Lipschitz continuous. For any given $r \ge 0$,*

$$\sup_{\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}} \le r} \|\boldsymbol{\Sigma}^{-1/2}\{\widehat{\mathbf{J}}_h(\boldsymbol{\delta}) - \mathbf{J}_h\}\boldsymbol{\Sigma}^{-1/2}\|_2 \lesssim \sqrt{\frac{p\log n + t}{nh}} + r \quad (3.32)$$

holds with probability at least $1 - e^{-t}$, provided that $\max\{\sqrt{(p+t)/n}, p\log(n)/n\} \lesssim h \lesssim 1$. The same probabilistic bound also applies to $\sup_{\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}} \le r} \|\boldsymbol{\Sigma}^{-1/2}\{\widehat{\mathbf{V}}_h(\boldsymbol{\delta}) - \mathbf{V}_h\}\boldsymbol{\Sigma}^{-1/2}\|_2$.

Following the discussions below Theorem 3.3.3, if we set the bandwidth as $h \asymp \{(p + \log n)/n\}^{2/5}$, $\|\widehat{\boldsymbol{\beta}}_h - \boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma}} = \mathscr{O}_{\mathbb{P}}(\sqrt{(p+\log n)/n})$ and $n^{1/2}\boldsymbol{a}^{\mathrm{T}}(\widehat{\boldsymbol{\beta}}_h - \boldsymbol{\beta}^*)/\sigma_h(\boldsymbol{a}) \to \mathscr{N}(0,1)$ in distribution uniformly over $\boldsymbol{a} \in \mathbb{R}^p$ as $n \to \infty$ under the constraint $p^{8/3} = o(n)$. With the same

bandwidth, it follows from Proposition 3.3.2 that

$$\max\left(\|\widehat{\mathbf{J}}_h - \mathbf{J}_h\|_2, \|\widehat{\mathbf{V}}_h - \mathbf{V}_h\|_2\right) = \mathscr{O}_{\mathbb{P}}\left[\left\{(\log n)^{1/2} p^{3/10} + (\log n)^{3/10} p^{1/2}\right\} n^{-3/10}\right] = o_{\mathbb{P}}(1).$$

This ensures the consistency of variance estimators, that is, $|\widehat{\sigma}_h^2(\boldsymbol{a})/\sigma_h^2(\boldsymbol{a}) - 1| \xrightarrow{\mathbb{P}} 0$.

3.4 Numerical Studies

In this section, we assess the finite-sample performance of conquer via extensive numerical studies. We compare conquer to standard QR [Koenker and Bassett, 1978] and Horowitz's smoothed QR [Horowitz, 1998]. Both the convolution-type and Horowitz's smoothed methods involve a smoothing parameter h. In view of Theorem 3.3.3, we take $h = \{(p + \log n)/n\}^{2/5}$ in all of the numerical experiments. In all the numerical experiments, the convergence criterion in Algorithms 1 and 2 is taken as $\delta = 10^{-4}$.

We first generate the covariates $\boldsymbol{x}_i = (x_{i,1}, \ldots, x_{i,p})^{\mathrm{T}}$ from a multivariate uniform distribution on the cube $3^{1/2} \cdot [-1,1]^p$ with covariance matrix $\boldsymbol{\Sigma} = (0.7^{|j-k|})_{1 \le j,k \le p}$ using the `R` package `MultiRNG` [Falk, 1999]. The random noise ε_i is generated from two different distributions: (i) Gaussian distribution, $\mathscr{N}(0,4)$; and (ii) t distribution with two degrees of freedom, t_2. Let $\boldsymbol{\beta}^* = (1, \ldots, 1)_p^{\mathrm{T}}$, and $\beta_0^* = 1$. Given $\tau \in (0,1)$, we then generate the response y_i from the following homogeneous and heterogeneous models:

1. Homogeneous model:

$$y_i = \beta_0^* + \langle \boldsymbol{x}_i, \boldsymbol{\beta}^* \rangle + \{\varepsilon_i - F_{\varepsilon_i}^{-1}(\tau)\}, \quad i = 1, \ldots, n; \tag{3.33}$$

2. Linear heterogeneous model:

$$y_i = \beta_0^* + \langle \boldsymbol{x}_i, \boldsymbol{\beta}^* \rangle + (0.5 x_{i,p} + 1)\{\varepsilon_i - F_{\varepsilon_i}^{-1}(\tau)\}, \quad i = 1, \ldots, n; \tag{3.34}$$

3. Quadratic heterogeneous model:

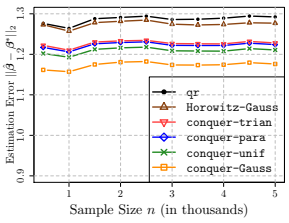
$$y_i = \beta_0^* + \langle \boldsymbol{x}_i, \boldsymbol{\beta}^* \rangle + 0.5\{1 + (x_{i,p} - 1)^2\}\{\varepsilon_i - F_{\varepsilon_i}^{-1}(\tau)\}, \quad i = 1, \ldots, n. \tag{3.35}$$

To evaluate the performance of different methods, we calculate the estimation error under the ℓ_2-norm, i.e., $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2$, and record the elapsed time. The details are in Section 3.4.1. In Section 3.4.2, we examine the finite-sample performance of the multiplier bootstrap method for constructing confidence intervals in terms of coverage probability, width of the interval, and computing time.

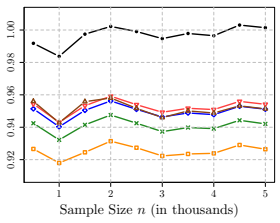
3.4.1 Estimation

For all the numerical studies in this section, we consider a wide range of the sample size n, with the size-dimension ratio fixed at $n/p = 20$. That is, we allow the dimension p to increase as a function of n. We implement conquer with four different kernel functions as described in Remark 3.2.1: (i) Gaussian; (ii) uniform; (iii) Epanechnikov; and (iv) triangular. The classical quantile regression is implemented via a modified version of the Barrodale and Roberts algorithm [Koenker and d'Orey, 1987, 1994] by setting `method=` "`br`" in the `R` package `quantreg`, which is recommended for problems with up to several thousands of observations in Koenker [2022]. For very large problems, the Frisch-Newton approach after preprocessing "pfn" is preferred. Since the same size taken to be at most 5000 throughout this section, the two methods, "br" and "pfn", have nearly identical runtime behaviors. In some applications where there are a lot of discrete covariates, it is advantageous to use method "sfn", a sparse version of Frisch-Newton algorithm that exploits sparse algebra to compute iterates [Koenker and Ng, 2003]. Moreover, we implement Horowitz's smoothed quantile regression using the Gaussian kernel, and solve the resulting non-convex optimization via gradient descent with random initialization and stepsize calibrated by backtracking line search (Section 9.3 of Boyd and Vandenberghe, 2004). The results, averaged over 500 replications, are reported.

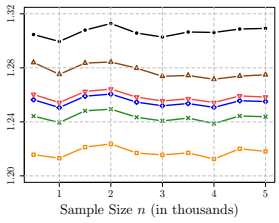
Figure 3.3 depicts estimation error of the different methods under the simulation settings described in Section 3.4 with $\tau = 0.9$. We see that conquer has a lower estimation error than the classical QR across all scenarios, indicating that smoothing can improve estimation accuracy under the finite-sample setting. Moreover, compared to Horowitz's smoothing, conquer has a lower estimation error in most settings. Estimation error under various quantile levels $\tau \in \{0.1, 0.3, 0.5, 0.7\}$ with the $\mathcal{N}(0,4)$ and t_2 random noise are also examined. The results are reported in He et al. [2022], from which we observe evident advantages of conquer, especially at low and high quantile levels.



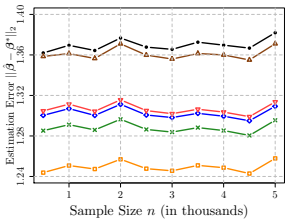
(a) Model (3.33) with $\mathcal{N}(0,4)$ error.



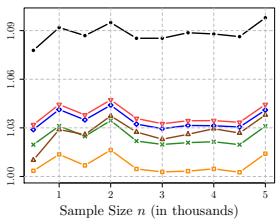
(b) Model (3.34) with $\mathcal{N}(0,4)$ error.



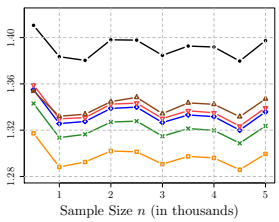
(c) Model (3.35) with $\mathcal{N}(0,4)$ error.



(d) Model (3.33) with t_2 error.



(e) Model (3.34) with t_2 error.

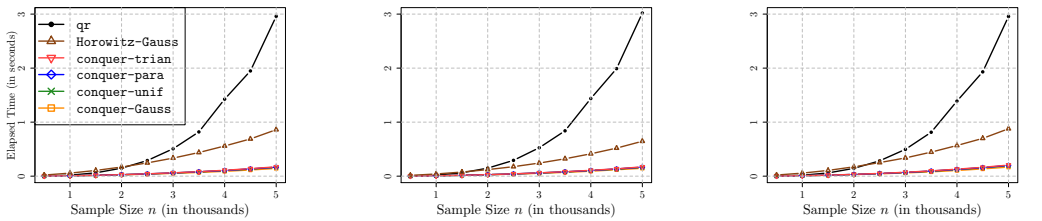


(f) Model (3.35) with t_2 error.

Figure 3.3. Estimation error under models (3.33)–(3.35) in Section 3.4 with $\mathcal{N}(0,4)$ and t_2 errors, $\tau = 0.9$, averaged over 500 data sets for three different methods: (i) quantile regression `qr`, (ii) Horowitz's method with Gaussian kernel `Horowitz-Gauss`, and (iii) the conquer method with four different kernel functions `conquer-trian`, `conquer-para`, `conquer-unif`, and `conquer-Gauss`.

To assess the computational efficiency, we compute the elapsed time for fitting the different methods. Figure 3.4 reports the runtime for the different methods with growing sample size and dimension under the same settings as in Figure 3.3. We observe that conquer is

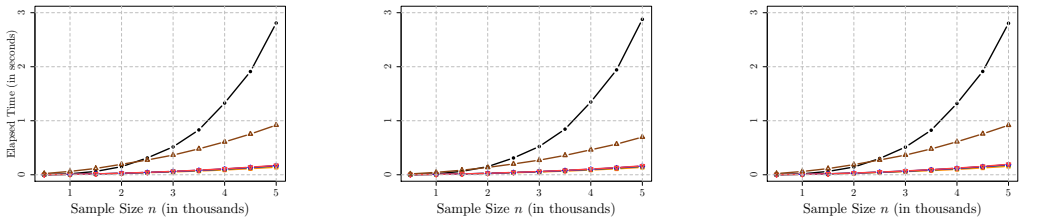
computationally efficient and stable across all scenarios, and the runtime is insensitive to the choice of kernel functions. In contrast, the runtime for classical quantile regression grows rapidly as the sample size and dimension increase. Figure 3.4 shows that the runtime of Horowitz's smoothing method increases significantly at extreme quantile levels $\tau \in \{0.1, 0.9\}$, possibly due to the combination of its non-convex nature and flatter gradient. In summary, we conclude that conquer significantly improves computational efficiency while retaining high statistical accuracy for fitting large-scale linear quantile regression models. Moreover, through a sensitivity analysis regarding the smoothing bandwidth h [He et al., 2022], it can be found that conquer is insensitive to the choice of bandwidth h.



(a) Model (3.33) with $\mathcal{N}(0,4)$ error.

(b) Model (3.34) with $\mathcal{N}(0,4)$ error.

(c) Model (3.35) with $\mathcal{N}(0,4)$ error.



(d) Model (3.33) with t_2 error.

(e) Model (3.34) with t_2 error.

(f) Model (3.35) with t_2 error.

Figure 3.4. Elapsed time of standard QR, Horowitz's smoothing, and conquer when $\tau = 0.9$. The model settings are the same as those in Figure 3.3.

3.4.2 Inference

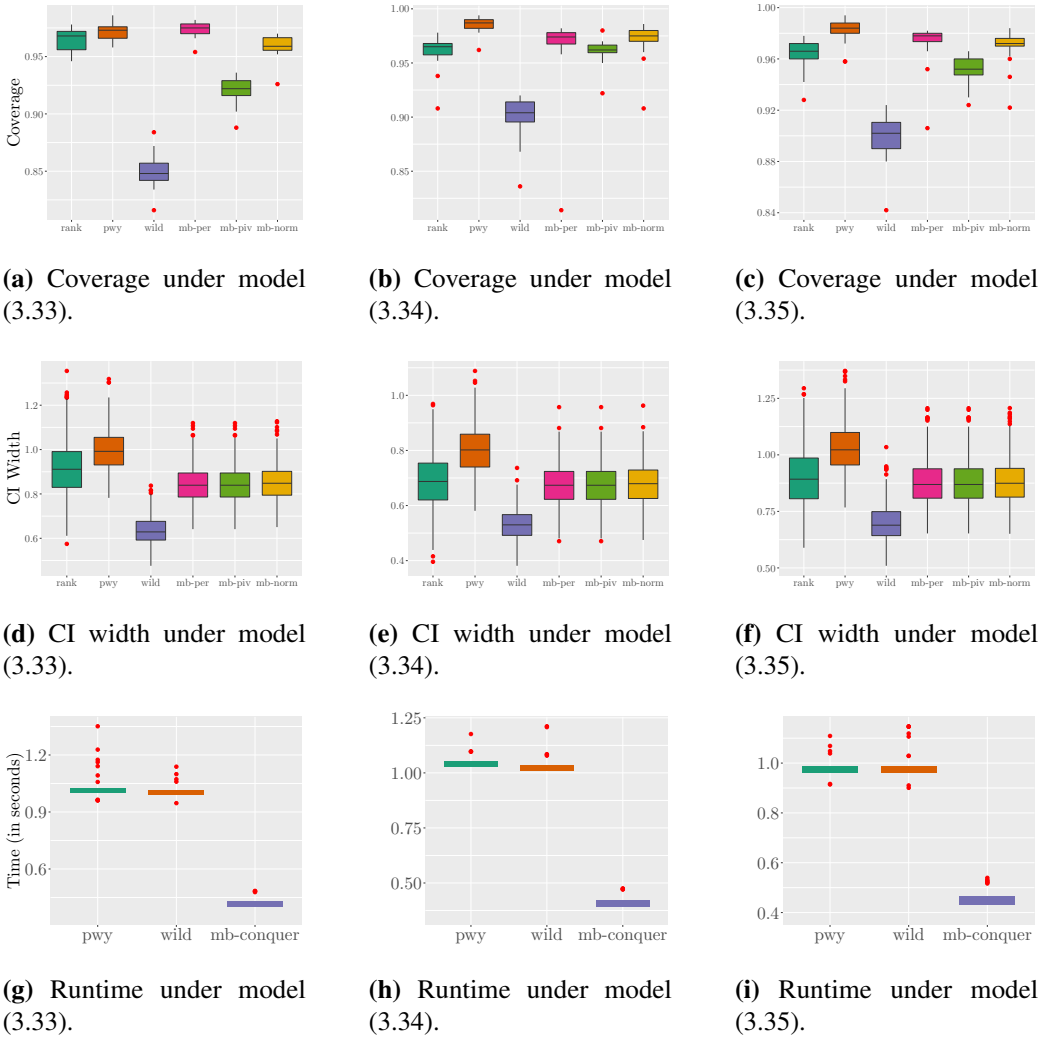
In this section, we assess the performance of the multiplier bootstrap procedure for constructing confidence interval for each of the regression coefficients obtained from conquer.

We implement conquer using the Gaussian kernel, and construct three types of confidence intervals: (i) the percentile `mb-per`; (ii) pivotal `mb-piv`; (iii) and regular `mb-norm` confidence intervals, as described in Section 3.1.3. We also refer to the proposed multiplier bootstrap procedure as `mb-conquer` for simplicity. We compare the proposed method to several widely used inference methods for QR. In particular, we consider confidence intervals by inverting a rank score test, `rank` (Gutenbrunner and Jurečková [1992]; Section 3.5 of Koenker [2005]); a bootstrap variant based on pivotal estimating functions, `pwy` [Parzen, Wei and Ying, 1994]; and wild bootstrap with Rademacher weights, `wild` [Feng, He and Hu, 2011]. The three methods `rank`, `pwy`, and `wild` are implemented using the `R` package `quantreg`. Note that `rank` is a non-resampling based procedure that relies on prior knowledge on the random noise, i.e., a user needs to specify whether the random noise are independent and identically distributed. In our simulation studies, we provide `rank` an unfair advantage by specifying the correct random noise structure.

We set $(n,p) = (800,20)$, $\tau \in \{0.5, 0.9\}$, and significance level $\alpha = 0.05$. All of the resampling methods are implemented with $B = 500$ bootstrap samples. To measure the reliability, accuracy, and computational efficiency of different methods for constructing confidence intervals, we calculate the average empirical coverage probability, average width of confidence interval, and the average runtime. The average is taken over all regression coefficients without the intercept. Results based on 500 replications are reported in Figure 3.5, and in He et al. [2022].

In these figures, we use the rank-inversion method, `rank`, as a benchmark since we implement `rank` using information about the true underlying random noise, which is practically infeasible. In the case of $\tau = 0.9$, `pwy` is the most conservative as it produces the widest confidence intervals with slightly inflated coverage probability, and `wild` gives the narrowest confidence intervals but at the cost of coverage probability. The proposed methods `mb-per`, `mb-piv`, and `mb-norm` achieve a good balance between reliability (high coverage probability) and accuracy (narrow CI width), and moreover, has the lowest runtime.

To further highlight the computational gain of the proposed method, we now perform



(a) Coverage under model (3.33).

(b) Coverage under model (3.34).

(c) Coverage under model (3.35).

(d) CI width under model (3.33).

(e) CI width under model (3.34).

(f) CI width under model (3.35).

(g) Runtime under model (3.33).

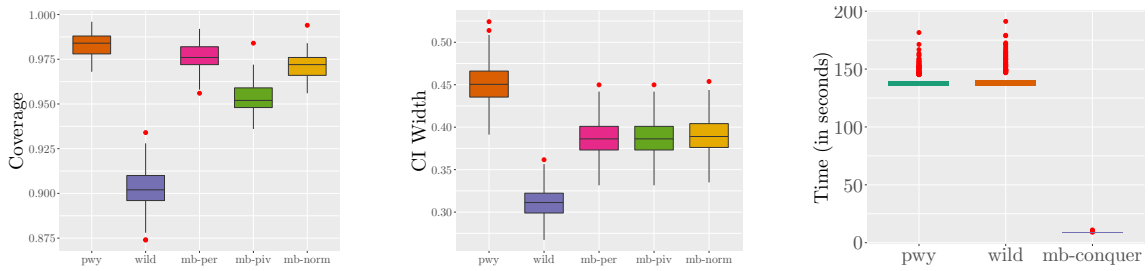
(h) Runtime under model (3.34).

(i) Runtime under model (3.35).

Figure 3.5. Empirical coverage, confidence interval width, and elapsed time of *six* methods: `rank`, `pwy`, `wild` and three types of mb-conquer: `mb-per`, `mb-piv`, and `mb-norm` under models (3.33)–(3.35) with t_2 errors. For the running time, `rank` is not included since it is not a resampling-based method. The quantile level τ is fixed to be 0.9, and the results are averaged over 500 data sets.

numerical studies with larger n and p. In this case, the rank inversion method `rank` is computationally infeasible. For example, when $(n, p) = (5000, 250)$, rank inversion takes approximately 80 minutes while conquer with multiplier bootstrap takes 41 seconds for constructing confidence intervals. We therefore omit `rank` from the following comparison. We consider the quadratic heterogeneous model (3.35) with $(n, p) = (4000, 100)$ and t_2 noise. The results are reported in Figure 3.6. We see that `pwy` and `wild` take up to 200 seconds while `mb-conquer` takes less than

10 seconds. In summary, `mb-conquer` leads to a huge computational gain without sacrificing statistical efficiency.



(a) Coverage **(b)** CI width **(c)** Runtime

Figure 3.6. Empirical coverage, confidence interval width and elapsed time of `pwy`, `wild` and 3 types of mb-conquer: `mb-per`, `mb-piv`, and `mb-norm` under quadratic heterogeneous model (3.35) with t_2 errors. This figure extends the rightmost column of Figure 3.5 to larger scale: $(n, p) = (4000, 100)$.

3.5 Discussion

In this chapter, we provide a comprehensive study on the statistical properties of *conquer*, namely, convolution-type smoothed quantile regression, under the non-asymptotic setting in which p is allowed to increase as a function of n while p/n being small. When a non-negative kernel is used, the smoothed objective function is convex, twice continuously differentiable, and locally strongly convex in a neighborhood of $\boldsymbol{\beta}^*$ (with high probability). An efficient gradient-based algorithm is proposed to compute the conquer estimator, which is scalable to very large-scale problems. For traditional QR computation with linear programming, interior point algorithms are typically used to get solutions with high precision (low duality gap) [Portnoy and Koenker, 1997]. When applied to large-scale datasets, this may be inefficient for two reasons: (i) it takes a lot more time to reach a duality gap of the order of machine precision, and (ii) such a generic algorithm, which is less tailored to problem structure, tends to be very slow or even run out of memory. In this regard, convolution smoothing offers a balanced tradeoff between statistical accuracy and computational complexity.

In the context of nonparametric density or regression estimation, it is known that when higher-order kernels are used (and if the density or regression function has enough derivatives), the bias is proportional to h^ν for some $\nu \geq 4$ which is of better order than h^2. Since a higher-order kernel has negative parts, the resulting smoothed loss is non-convex and thus brings the computational issue once again. Motivated by the two-stage procedure proposed by Bickel [1975] whose original idea is to improve an initial estimator that is already consistent but not efficient, we further propose a one-step conquer estimator using higher-order kernels but without the need for solving a large-scale non-convex optimization. With increasing degrees of smoothness, the one-step conquer is asymptotically normal under a milder dimension constraint of roughly $p^2/n \to 0$. Due to space limitations, the details of this method are relegated to Section B.1 in the supplementary material.

In high-dimensional settings in which $p \gg n$, various authors have studied the regularized quantile regression under the sparsity assumption that most of the regression coefficients are zero [Belloni and Chernozhukov, 2011, Wang, Wu and Li, 2012, Zheng, Peng and He, 2015]. The computation of ℓ_1-penalized QR is based on either reformulation as linear programs or alternating direction method of multiplier algorithms [Gu et al., 2018]. The theory and computation of regularized conquer with flexible penalties have been developed in Tan, Wang and Zhou [2021] and Man et al. [2022]. In Chapter 4, we analyze regularized conquer in a more complicated scenario with randomly censored outcomes, and show that gradient-based algorithms enjoy superior computational efficiency without sacrificing statistical accuracy.

3.6 Acknowledgements

This chapter, in part, is a reprint of the material in the paper "Smoothed quantile regression with large-scale inference", He, Xuming; Pan, Xiaoou; Tan, Kean Ming and Zhou, Wen-Xin. The paper has been accepted by *Journal of Econometrics*, 2022. The dissertation author was the primary investigator and author of this paper.

Chapter 4

Scalable Learning on Censored Data

4.1 Overview

In this chapter, we develop a smoothed framework for censored quantile regression (CQR) that is scalable to problems with large dimension p in both the low- and high-dimensional settings. Our proposed method is motivated by the smoothed estimating equation approach that has surfaced mostly in the econometrics literature [Whang, 2006, Wu, Ma and Yin, 2015, Kaplan and Sun, 2017, de Castro et al., 2019, Fernandes, Guerre and Horta, 2021, He et al., 2022], which can be applied to the stochastic integral based sequential estimation procedure proposed by Peng and Huang [2008] for CQR. We show in Section 4.2.2 that the smoothed sequential estimating equations method can be reformulated as solving a sequence of optimization problems with (at least) twice-differentialble and convex loss functions for which gradient-based algorithms are available. Large-scale statistical inference can then be performed efficiently via multiplier/weighted bootstrap. In the high-dimensional setting, we propose and analyze ℓ_1-penalized smoothed CQR estimators obtained by sequentially minimizing smoothed convex loss functions plus ℓ_1-penalty, which we solve using a scalable and efficient majorize-minimization-type algorithm.

This chapter is mainly motivated by the computational challenge for CQR. To illustrate this issue, we compare the ℓ_1-penalized CQR proposed by Zheng, Peng and He [2018] and our proposed method by analyzing a gene expression dataset studied in Shedden et al. [2008]. In this

study, 22,283 genes from 442 lung adenocarcinomas are incorporated to predict the survival time in lung cancer, with 46.6% subjects that are censored. We implement both methods with quantile levels grid set as $\{0.1, 0.11, \ldots, 0.7\}$, and use a predetermined sequence of penalty weights. For Zheng, Peng and He [2018], we use the `rqPen` package to compute the ℓ_1-penalized QR estimator at each quantile level [Sherwood and Maidman, 2022]. The computational time and maximum allocated memory are reported in Table 4.1. The reference machine for this experiment is a worker node with 2.5 GHz 32-core processor and 512 GB of memory in a high-performance computing cluster.

Table 4.1. Computational runtime and maximum allocated memory for fitting ℓ_1-penalized CQR and the proposed method on the gene expression data with censored response in Shedden et al. [2008]. One gigabyte (GB) equals 1024 megabytes (MB).

Methods	Runtime	Allocated memory
ℓ_1-penalized CQR	170 hours+	38 GB
Proposed method	2 minutes	926 MB

Theoretically, we provide a unified analysis for the proposed smoothed estimator in both low- and high-dimensional settings. In the low-dimensional case where the dimension is allowed to increase with the sample size, we establish the uniform rate of convergence and a uniform Bahadur-type representation for the smoothed CQR estimator. We also provide a rigorous justification for the validity of a weighted/multiplier bootstrap procedure with explicit error bounds as functions of (n, p). To our knowledge, these are the first results for censored quantile regression in the increasing-p regime with $p < n$. The main challenges are as follows. To fit the QR process with censored response variables, the stochastic integral based approach entails a sequence of estimating equations which correspond to a prespecified grid of quantile indexes. A sequence of pointwise estimators can then be sequentially obtained by solving these equations. The recursive nature of this procedure poses technical challenges because at each quantile level, the objective function (or the estimating equation) depends on all of the previous estimates. To establish convergence rates for the estimated regression process, a delicate analysis

beyond what is used in He et al. [2022] is required to deal with the accumulated estimation error sequentially. The mesh width of the grid should converge to zero at a proper rate in order to balance the accumulated estimation error and discretization error. In the high-dimensional setting, we show that with suitably chosen penalty levels and bandwidth, the ℓ_1-penalized smoothed CQR estimator has a uniform convergence rate of $\mathcal{O}(\sqrt{s\log(p)/n})$, provided the sample size satisfies $n \gtrsim s^3\log(p)$. The technical arguments used in this case are also very different from those in Zheng, Peng and He [2018] and subsequent work Fei et al. [2021], and as a result, our conclusion improves that of Zheng, Peng and He [2018] by relaxing the exponential term $\exp(Cs)$ in the convergence rate to a linear term in s. Such an improvement is significant when the effective model size s is allowed to grow with n and p in the context of censored quantile regression.

4.2 Censored Quantile Regression

Let $z \in \mathbb{R}$ be a response variable of interest, and $\boldsymbol{x} = (x_1, \ldots, x_p)^{\mathrm{T}}$ be a p-vector ($p \geq 2$) of random covariates with $x_1 \equiv 1$. In this work, we focus on a global conditional quantile model on z described as follows. Given a closed interval $[\tau_L, \tau_U] \subseteq (0,1)$, assume that the τ-th conditional quantile of z given $\boldsymbol{x}$ takes the form

$$F^{-1}_{z|\boldsymbol{x}}(\tau) = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*(\tau) \ \text{ for any } \ \tau \in [\tau_L, \tau_U], \tag{4.1}$$

where $\boldsymbol{\beta}^*(\tau) \in \mathbb{R}^p$, formulated as a function of τ, is the unknown vector of regression coefficients.

We assume that z is subject to right censoring by C, a random variable that is conditionally independent of z given the covariates $\boldsymbol{x}$. Let $y = z \wedge C$ the censored outcome, and $\Delta = \mathbb{1}(z \leq C)$ be an event indicator. The observed samples $\{y_i, \Delta_i, \boldsymbol{x}_i\}_{i=1}^n$ consist of independent and identically distributed (i.i.d.) replicates of the triplet $(y, \Delta, \boldsymbol{x})$. In addition, we assume at the outset that the lowest quantile of interest τ_L satisfies $\mathbb{P}\{y \leq \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*(\tau_L), \Delta = 0\} = 0$. This condition, interpreted as no censoring below the τ_L-th quantile, is commonly imposed in the context of CQR; see, e.g., Condition C in Portnoy [2003] and Assumption 3.1 in Zheng, Peng and He [2018]. Moreover, our

quantiles of interest are confined up to $\tau_U < 1$ subject to some identifiability concerns, which is a subtle issue for CQR problems. Briefly speaking, the model (4.1) may become non-identifiable as τ moves towards 1, due to large amount of censored information in the upper tail. In practice, determining τ_U is usually a compromise between inference range of interest and data censoring rate, and determining τ_L requires a careful investigation if censoring can occur at early stages. Theoretically, the above assumption on τ_L helps us simplify the technical arguments.

The above model is broadly defined, yet it is inspired by approaching survival data with quantile regression [Koenker and Geling, 2001]. To briefly illustrate, let T be a non-negative random variable representing the failure time to an event. The conditional quantile model (4.1) on $z = \log(T)$ can be viewed as a generalization of the standard accelerated failure time model in the sense that coefficients not only shift the location but also affect the shape and dispersion of the conditional distributions.

4.2.1 Martingale-based estimating equation estimator

Under the global linear model (4.1), two well-known methods are the recursively re-weighted estimator of Portnoy [2003] and the stochastic integral based estimating equation estimator of Peng and Huang [2008]. Both methods are grid-based algorithms that iteratively solve a sequence of (weighted) check function minimization problems over a predetermined grid of τ-values. Motivated by the recent success of smoothing methods for uncensored parametric and nonparametric quantile regressions [Fasiolo et al., 2021, Fernandes, Guerre and Horta, 2021, He et al., 2022], we propose a smoothed estimating equation approach for CQR in the next subsection. We start with a brief introduction of Peng and Huang [2008]'s method that is built upon the martingale structure of randomly censored data.

To this end, denote $\Lambda_{z|\boldsymbol{x}}(t) = -\log\{1 - \mathbb{P}(z \le t|\boldsymbol{x})\}$ as the cumulative conditional hazard function of z given $\boldsymbol{x}$, and define the counting processes as $N_i(t) = \mathbb{1}\{y_i \le t, \Delta_i = 1\}$ and $N_{0i}(t) = \mathbb{1}\{y_i \le t, \Delta_i = 0\}$ for $i = 1, \ldots, n$, where $\Delta_i = \mathbb{1}(z_i \le C_i)$. Define $\mathscr{F}_i(s) = \sigma\{N_i(u), N_{0i}(u) : u \le s\}$ as the σ-algebra generated by the foregoing processes. Note that $\{\mathscr{F}_i(s) : s \in \mathbb{R}\}$ is an increasing

family of sub-σ-algebras, also known as filtration, and $N_i(t)$ is an adapted sub-martingale. By the unique Doob-Meyer decomposition, one can construct an $\mathscr{F}_i(t)$-martingale $M_i(t) = N_i(t) - \Lambda_{z|\boldsymbol{x}_i}(y_i \wedge t)$ satisfying $\mathbb{E}\{M_i(t)|\boldsymbol{x}_i\} = 0$; see Section 1.3 of Fleming and Harrington [1991] for details. Taking $t = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*(\tau)$ for each i, the martingale property implies

$$\mathbb{E}\left[\sum_{i=1}^{n}\left\{N_i\big(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*(\tau)\big) - \Lambda_{z|\boldsymbol{x}_i}\big(y_i \wedge \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*(\tau)\big)\right\}\boldsymbol{x}_i\right] = \mathbf{0}.$$

This lays the foundation for the stochastic integral based estimating equation approach. The monotonicity of the function $\tau \mapsto \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*(\tau)$, implied by the global linearity in (4.1), leads to

$$\Lambda_{z|\boldsymbol{x}_i}\big(y_i \wedge \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*(\tau)\big) = H(\tau) \wedge H\big(\mathbb{P}(z \le y_i|\boldsymbol{x}_i)\big) = \int_0^{\tau} \mathbb{1}\{y_i \ge \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*(u)\}\,\mathrm{d}H(u)$$

for $\tau \in [\tau_L, \tau_U]$, where $H(u) := -\log(1-u)$ for $0 < u < 1$. This motivates Peng and Huang's estimator [Peng and Huang, 2008], which solves the following estimating equation

$$\frac{1}{n}\sum_{i=1}^{n}\left[N_i\big(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}(\tau)\big) - \int_0^{\tau} \mathbb{1}\{y_i \ge \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}(u)\}\,\mathrm{d}H(u)\right]\boldsymbol{x}_i = \mathbf{0}, \text{ for every } \tau_L \le \tau \le \tau_U.$$

However, the exact solution to the above equation is not directly obtainable. By adapting Euler's forward method for ordinary differential equation, Peng and Huang [2008] proposed a grid-based sequential estimating procedure as follows. Let $\tau_L = \tau_0 < \tau_1 < \cdots < \tau_m = \tau_U$ be a grid of quantile indices. Noting that $\mathbb{P}\{y \le \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*(\tau_0), \Delta = 0\} = 0$, we have $\mathbb{E}\int_0^{\tau_0} \mathbb{1}\{y_i \ge \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*(u)\}\,\mathrm{d}H(u) = \tau_0$, and hence $\boldsymbol{\beta}^*(\tau_0)$ can be estimated by solving the usual quantile equation $(1/n)\sum_{i=1}^{n}\{N_i(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}) - \tau_0\}\boldsymbol{x}_i = \mathbf{0}$. Denote $\widetilde{\boldsymbol{\beta}}(\tau_0)$ as the solution to the above equation. At grid points τ_k, $k = 1, \ldots, m$, the estimators $\widetilde{\boldsymbol{\beta}}(\tau_k)$ are sequentially obtained by solving

$$\frac{1}{n}\sum_{i=1}^{n}\left[N_i(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}) - \sum_{j=0}^{k-1}\int_{\tau_j}^{\tau_{j+1}} \mathbb{1}\{y_i \ge \boldsymbol{x}_i^{\mathrm{T}}\widetilde{\boldsymbol{\beta}}(\tau_j)\}\,\mathrm{d}H(u) - \tau_0\right]\boldsymbol{x}_i = \mathbf{0}. \tag{4.2}$$

The resulting estimated function $\widetilde{\boldsymbol{\beta}}(\cdot) : [\tau_L, \tau_U] \mapsto \mathbb{R}^p$ is right-continuous and piecewise-constant

that jumps only at each grid point. Computationally, solving the above equation is equivalent to minimizing an ℓ_1-type convex objective function after introducing a sufficiently large pseudo point. The minimizer, however, is not always uniquely defined. To avoid this lack of uniqueness as well as grid dependence, Huang [2010] introduced a more general (population) integral equation, and then proposed a Progressive Localized Minimization (PLMIN) algorithm solve its empirical version exactly. This algorithm automatically determines the breakpoints of the solution and thus is grid-free. Under a continuity condition on the density functions (see, e.g. condition (C2) in Huang [2010]), the estimating functions used in Peng and Huang [2008] and Huang [2010] are asymptotically equivalent.

4.2.2 A smoothed estimating equation approach

Due to the discontinuity stemming from the indicator function in the counting process $N_i(\cdot)$, exact solutions to the estimating equations (4.2) may not exist. In fact, $\widetilde{\boldsymbol{\beta}}(\tau_j)$ for $j = 0,\ldots,m$ are defined as the general solutions to generalized estimating equations [Fygenson and Ritov, 1994], which correspond to subgradients of some convex yet non-differentiable functions. Computationally, one may reformulate these equations as a sequence of linear programs, solvable by the Frisch-Newton algorithm described in Portnoy and Koenker [1997]. The computation complexity grows rapidly when the dimensionality p increases with the sample size. To mitigate the computational burden of the existing methods, we use a smoothed estimating equation (SEE) approach for fitting large-scale censored quantile regression models.

Let $K(\cdot)$ be a symmetric and non-negative kernel function and let $\bar{K}(u) = \int_{-\infty}^{u} K(x)\,\mathrm{d}x$, which is a non-decreasing function that is between 0 and 1. The non-smooth indicator function $\mathbb{1}(u \geq 0)$ can thus be approximated by $\bar{K}(u/h)$ for some $h > 0$ in the sense that as $h \to 0$, $\bar{K}(u/h) \to 1$ for $u \geq 0$ and $\bar{K}(u/h) \to 0$ for $u < 0$. Hereinafter, $h > 0$ will be referred to as a bandwidth. As the aforementioned, let $\tau_L = \tau_0 < \tau_1 < \cdots < \tau_m = \tau_U$ be a grid of quantile indices

for some $m \geq 1$. Given a kernel function $K(\cdot)$ and a bandwidth $h > 0$, write

$$K_h(u) = h^{-1}K(u/h) \quad \text{and} \quad \bar{K}_h(u) = \bar{K}(u/h) = \int_{-\infty}^{u/h} K(v)\mathrm{d}v, \;\; u \in \mathbb{R},$$

so that $\bar{K}_h'(u) = K_h(u)$. We now propose a smooth SEE approach for CQR.

1. At $\tau = \tau_0$, we estimate $\boldsymbol{\beta}^*(\tau_0)$ by $\widehat{\boldsymbol{\beta}}(\tau_0)$, obtained from solving $\widehat{Q}_0(\boldsymbol{\beta}) = \mathbf{0}$, where

$$\widehat{Q}_0(\boldsymbol{\beta}) := \frac{1}{n}\sum_{i=1}^{n} \big\{\Delta_i \bar{K}_h(-r_i(\boldsymbol{\beta})) - \tau_0\big\}\boldsymbol{x}_i \quad \text{and} \quad r_i(\boldsymbol{\beta}) = y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}. \tag{4.3}$$

2. At grid points τ_k for $k = 1, \ldots, m$, set $\widehat{\boldsymbol{\beta}}(\tau) = \widehat{\boldsymbol{\beta}}(\tau_{k-1})$ for any $\tau \in (\tau_{k-1}, \tau_k)$, and then obtain estimators $\widehat{\boldsymbol{\beta}}(\tau_k)$ of $\boldsymbol{\beta}^*(\tau_k)$ by solving $\widehat{Q}_k(\boldsymbol{\beta}) = \mathbf{0}$, where

$$\widehat{Q}_k(\boldsymbol{\beta}) := \frac{1}{n}\sum_{i=1}^{n} \left[\Delta_i \bar{K}_h(-r_i(\boldsymbol{\beta})) - \sum_{j=0}^{k-1} \bar{K}_h(r_i(\widehat{\boldsymbol{\beta}}(\tau_j))\{H(\tau_{j+1}) - H(\tau_j)\} - \tau_0\right]\boldsymbol{x}_i. \tag{4.4}$$

Note that the resulting estimator $\widehat{\boldsymbol{\beta}}(\cdot) : [\tau_L, \tau_U] \mapsto \mathbb{R}^p$ is right-continuous and piecewise-constant with jumps only at grids. For notational convenience, throughout the manuscript, let

$$\boldsymbol{\beta}_k^* = \boldsymbol{\beta}^*(\tau_k) \quad \text{and} \quad \widehat{\boldsymbol{\beta}}_k = \widehat{\boldsymbol{\beta}}(\tau_k), \quad k = 0, 1, \ldots, m.$$

Before proceeding, it is worth noticing that the above smoothed estimating equations method is closely related to the convolution smoothing approach studied in Chapter 3. Consider the check function $\rho_\tau(u) = \tau\{u - \mathbb{1}(u < 0)\}$, and its convolution smoothed counterpart

$$\ell_{\tau,h}(u) = (\rho_\tau * K_h)(u) = \int_{-\infty}^{\infty} \rho_\tau(v) K_h(v - u)\,\mathrm{d}v,$$

where $*$ denotes the convolution operator. Given censored data $\{(y_i, \Delta_i, \boldsymbol{x}_i)\}_{i=1}^n$, define the

empirical smoothed loss

$$\widehat{L}_0(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\left\{\Delta_i \ell_{\tau_0,h}(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}) + \tau_0(\Delta_i - 1)\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right\}, \tag{4.5}$$

whose gradient and Hessian are

$$\nabla\widehat{L}_0(\boldsymbol{\beta}) = \widehat{Q}_0(\boldsymbol{\beta}) \quad \text{and} \quad \nabla^2\widehat{L}_0(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\Delta_i K_h(r_i(\boldsymbol{\beta}))\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}},$$

respectively. Hence, the foregoing estimator $\widehat{\boldsymbol{\beta}}_0$ can be equivalently defined as the solution to the (unconstrained) optimization problem $\min_{\boldsymbol{\beta}\in\mathbb{R}^p}\widehat{L}_0(\boldsymbol{\beta})$. When a non-negative kernel is used, the objective function $\widehat{L}_0(\cdot)$ is convex, and thus any minimizer satisfies the first-order condition. At subsequent grid points τ_k for $k = 1,\ldots,m$, the estimator $\widehat{\boldsymbol{\beta}}_k$ can also be viewed as an M-estimator that solves

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p}\left\{\widehat{L}_k(\boldsymbol{\beta}) := \widehat{L}_0(\boldsymbol{\beta}) - \left\langle \frac{1}{n}\sum_{i=1}^{n}\sum_{j=0}^{k-1}\bar{K}_h(y_i - \boldsymbol{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_j)\{H(\tau_{j+1}) - H(\tau_j)\}\boldsymbol{x}_i, \boldsymbol{\beta}\right\rangle\right\}. \tag{4.6}$$

Notably, kernel smoothing produces continuously differentiable estimating functions $\widehat{Q}_k(\cdot)$ ($k = 0,\ldots,m$), or equivalently, convex and twice-differentiable loss functions $\widehat{L}_k(\cdot)$, which have the same positive semi-definite Hessian matrix $\nabla^2\widehat{L}_k(\boldsymbol{\beta}) = (1/n)\sum_{i=1}^{n}\Delta_i K_h(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} - y_i)\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}$. As we shall see, the empirical loss functions $\widehat{L}_k(\cdot)$ are not only globally convex but also locally strongly convex (with high probability). This property ensures the existence of global solutions to the sequential estimation problems, which can efficiently solved by a quasi-Newton algorithm.

4.2.3 Inference with bootstrapped process

In this subsection, we construct component-wise confidence intervals for $\boldsymbol{\beta}^*(\tau)$ at some quantile index τ of interest by bootstrapping the quantile process, following the idea in Section 2.1.2. Recall that $\widehat{\boldsymbol{\beta}}_k$'s are the solutions to the equations $\widehat{Q}_k(\boldsymbol{\beta}) = \mathbf{0}$, where $\widehat{Q}_k(\cdot)$

($k = 0, 1, \ldots, m$) are defined in (4.3) and (4.4). Analogously, we construct bootstrap estimators $\widehat{\boldsymbol{\beta}}_k^{\flat}$ following a sequential procedure based on the bootstrapped SEEs obtained by perturbing $\widehat{Q}_k(\cdot)$ with random weights. Independent of the observed data $\{y_i, \Delta_i, \boldsymbol{x}_i\}_{i=1}^n$, let $W_1, \ldots, W_n$ be exchangeable non-negative random variables, satisfying $\mathbb{E}(W_i) = 1$ and $\mathrm{var}(W_i) > 0$. The bootstrap estimators can be constructed as follows:

1. Set $\widehat{\boldsymbol{\beta}}_0^{\flat}$ as the solution of $\widehat{Q}_0^{\flat}(\boldsymbol{\beta}) = \mathbf{0}$, where

$$\widehat{Q}_0^{\flat}(\boldsymbol{\beta}) := \frac{1}{n}\sum_{i=1}^{n} W_i\{\Delta_i \bar{K}_h(-r_i(\boldsymbol{\beta})) - \tau_0\}\boldsymbol{x}_i \ \text{ with } \ r_i(\boldsymbol{\beta}) = y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}. \tag{4.7}$$

2. For $k = 1, \ldots, m$, compute $\widehat{\boldsymbol{\beta}}_k^{\flat}$ sequentially by solving $\widehat{Q}_k^{\flat}(\boldsymbol{\beta}) = \mathbf{0}$, where

$$\widehat{Q}_k^{\flat}(\boldsymbol{\beta}) := \frac{1}{n}\sum_{i=1}^{n} W_i\left[\Delta_i \bar{K}_h(-r_i(\boldsymbol{\beta})) - \sum_{\ell=0}^{k-1} \bar{K}_h(r_i(\widehat{\boldsymbol{\beta}}_\ell^{\flat}))\{H(\tau_{\ell+1} - H(\tau_\ell)\} - \tau_0\right]\boldsymbol{x}_i. \tag{4.8}$$

3. Define the bootstrap estimate of the coefficient process $\widehat{\boldsymbol{\beta}}^{\flat}(\cdot) : [\tau_L, \tau_U] \mapsto \mathbb{R}^p$ as $\widehat{\boldsymbol{\beta}}^{\flat}(\tau) = \widehat{\boldsymbol{\beta}}_{k-1}^{\flat}$ for $\tau \in [\tau_{k-1}, \tau_k)$ and $k = 1, \ldots, m$.

For a prescribed nominal level, we can construct component-wise percentile or normal-based confidence intervals for $\boldsymbol{\beta}_j^*(\tau)$ ($j = 1, \ldots, p$). The above multiplier bootstrap estimator $\widehat{\boldsymbol{\beta}}^{\flat}(\cdot) : [\tau_L, \tau_U] \mapsto \mathbb{R}^p$ of the coefficient process behaves similarly as $\widehat{\boldsymbol{\beta}}(\cdot)$, in the sense that they are both right-continuous and piecewise-constant with jumps only at the grids.

We complete this section with a brief discussion of other resampling methods for quantile regression. Given the random weights $\{W_i\}_{i=1}^n$ independent of data, another available approach is to minimize the randomly perturbed objective functions [Jin, Ying and Wei, 2001, Peng and Huang, 2008]. In the current setting, it seems more natural to directly bootstrap the estimating equations. In terms of bootstrapping estimating equations with uncensored data, Parzen, Wei and Ying [1994]'s method is based on the assumption that the estimating equation is exactly or

asymptotically pivotal, and Hu and Kalbfleisch [2000]'s proposal is based on resampling with replacement. A generalized weighted bootstrap and its asymptotic theory has been rigorously studied in Chatterjee and Bose [2005] and Ma and Kosorok [2005]. For censored quantile regression, the sequential SEEs (4.4) are not directly formulated as empirical averages of independent random quantities, nor do they satisfy the required assumptions in the literature; see Section 2 of Parzen, Wei and Ying [1994], Section 2 of Hu and Kalbfleisch [2000], and Section 3 of Chatterjee and Bose [2005]. Hence, the validity of weighted bootstrap for CQR is of independent interest, and will be examined in Section 4.3.3.

Remark 4.2.1. *In practice, random weights* $\{W_i\}_{i=1}^n$ *can be generated from one of the following distributions. (i)* $(W_1,\ldots,W_n) \sim \text{Multinomial}(n, 1/n, \ldots, 1/n)$. *This leads to Efron's nonparametric bootstrap, for which the random weights are exchangeable but not independent; (ii)* $W_1,\ldots,W_n \sim \text{Exp}(1)$ *are i.i.d. exponentially distributed random variables; and (iii)* $W_i = e_i + 1$, *where* e_i*'s are i.i.d. Rademacher random variables, defined by* $\mathbb{P}(e_i = 1) = \mathbb{P}(e_i = 0) = 1/2$. *We refer to this as the Rademacher multiplier bootstrap. Its theoretical properties will be investigated in Section 4.3.3.*

4.3 Theoretical Analysis

4.3.1 Regularity conditions

We first impose some technical assumptions required for the results in Sections 4.3.2 and 4.3.3.

Condition 4.3.1 (Kernel function)**.** *Let* $K(\cdot)$ *be a symmetric, Lipschitz continuous and non-negative kernel function, that is,* $K(u) = K(-u)$, $K(u) \geq 0$ *for all* $u \in \mathbb{R}$ *and* $\int_{-\infty}^{\infty} K(u)\,\mathrm{d}u = 1$. *Moreover,* $\kappa_u = \sup_{u\in\mathbb{R}} K(u) < \infty$, $\kappa_l = \min_{|u|\leq c} K(u) > 0$ *for some* $c > 0$. *We define its higher-order absolute moments as* $\kappa_\ell = \int_{-\infty}^{\infty} |u|^\ell K(u)\mathrm{d}u$ *for any positive integer* ℓ.

Condition 4.3.2 (Random design)**.** *The random covariate vector* $\boldsymbol{x} = (x_1,\ldots,x_p)^{\mathrm{T}} \in \mathscr{X} \subseteq \mathbb{R}^p$ *is compactly supported with* $\zeta_p := \sup_{\boldsymbol{x}\in\mathscr{X}} \|\Sigma^{-1/2}\boldsymbol{x}\|_2 < \infty$, *where* $\Sigma = \mathbb{E}(\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}})$ *is positive definite.*

Condition 4.3.3 (Conditional densities)**.** *Assume* $(z,\boldsymbol{x})$ *follows the global conditional quantile model* (4.1)*. Define the conditional cumulative distribution functions* $F_z(u|\boldsymbol{x}) = \mathbb{P}(z \le u|\boldsymbol{x})$*,* $F_y(u|\boldsymbol{x}) = \mathbb{P}(y \le u|\boldsymbol{x})$ *and* $G(u|\boldsymbol{x}) = \mathbb{P}(y \le u, \Delta = 1|\boldsymbol{x})$*, where* $y = z \wedge C$ *and* C *is independent of* z *given* $\boldsymbol{x}$*. Assume that the conditional densities* $f_z(u|\boldsymbol{x}) = F_z'(u|\boldsymbol{x})$*,* $f_y(u|\boldsymbol{x}) = F_y'(u|\boldsymbol{x})$ *and* $g(u|\boldsymbol{x}) = G'(u|\boldsymbol{x})$ *exist, and satisfy almost surely (over* $\boldsymbol{x}$*) that*

$$\inf_{\tau\in[\tau_L,\tau_U]} \min\left\{f_y(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*(\tau)|\boldsymbol{x}), f_z(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*(\tau)|\boldsymbol{x})\right\} \ge \underline{f} > 0, \quad \sup_{u\in\mathbb{R}} f_y(u|\boldsymbol{x}) \le \overline{f},$$

$$0 < \underline{g} \le \inf_{|u-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*(\tau)|\le 1/2, \tau\in[\tau_L,\tau_U]} g(u|\boldsymbol{x}) \le \sup_{u\in\mathbb{R}} g(u|\boldsymbol{x}) \le \overline{g}.$$

Moreover, there exists a constant $l_1 > 0$ *such that for any* $u \in \mathbb{R}$*,*

$$\sup_{\boldsymbol{x}\in\mathbb{R}^p, \tau\in[\tau_L,\tau_U]} |f_y(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*(\tau)+u|\boldsymbol{x}) - f_y(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*(\tau)|\boldsymbol{x})| \le l_1|u|,$$

$$\sup_{\boldsymbol{x}\in\mathbb{R}^p, \tau\in[\tau_L,\tau_U]} |g(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*(\tau)+u|\boldsymbol{x}) - g(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*(\tau)|\boldsymbol{x})| \le l_1|u|.$$

Condition 4.3.4 (Grid size)**.** *The grid of quantile levels* $\tau_L = \tau_0 < \tau_1 < \cdots < \tau_m = \tau_U$ *satisfies* $n^{-1} \le \delta_* \le \delta^* \lesssim n^{-1/2}$*, where* $\delta^* = \max_{1\le k\le m}(\tau_k - \tau_{k-1})$ *and* $\delta_* = \min_{1\le k\le m}(\tau_k - \tau_{k-1})$*.*

Condition 4.3.1 is similar to Condition 3.3.1. To simplify the analysis, we take $c = 1$ in Condition 4.3.1; otherwise if $c < 1$ and $K(\pm 1) = 0$, we can simply use a re-scaled kernel $K_c(u) := cK(cu)$, so that $\min_{|u|\le 1} K_c(u) = c\min_{|u|\le c} K(u)$. The compactness of $\mathscr{X}$ in Condition 4.3.2 is a common requirement for a global linear quantile regression model (quantile regression process) [Koenker, 2005]. If the support of the covariate space—the set of x_j's that occur with positive probability—is unbounded, at some points there will be "crossings" of the conditional quantile functions, unless these functions are parallel, which corresponds to a pure location-shift model. The quantity ζ_p plays an important role in the theoretical results. Alternatively, one may assume $\|\Sigma^{-1/2}\boldsymbol{x}\|_\infty \le C_0$ (almost surely) as in Zheng, Peng and He [2018], which in turn implies $\zeta_p \le C_0 p^{1/2}$ in the worst-case scenario. In general, it is reasonable to assume that $\zeta_p \asymp p^{1/2}$. In

addition to ζ_p, define the moment parameters

$$m_q = \sup_{\boldsymbol{u}\in\mathbb{S}^{p-1}} \mathbb{E}\big(|\boldsymbol{u}^{\mathrm{T}}\Sigma^{-1/2}\boldsymbol{x}|^q\big) \quad \text{for} \quad q = 3,4, \tag{4.9}$$

which satisfy the worst-case bounds $m_3 \le \zeta_p$ and $m_4 \le \zeta_p^2$.

Conditions 4.3.2 and 4.3.3 ensure that the coefficient function $\boldsymbol{\beta}^*(\cdot)$ is Lipschitz continuous. Since $\boldsymbol{\beta}^*(\tau)$ solves the equation $\mathbb{E}[\{\tau - \mathbb{1}(z \le \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta})\}\boldsymbol{x}] = \mathbf{0}$, we have $\frac{d}{d\tau}\boldsymbol{\beta}^*(\tau) = \mathbb{E}\{f_z(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*(\tau)|\boldsymbol{x})\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\}^{-1}\mathbb{E}(\boldsymbol{x})$. Under Condition 4.3.2, it holds

$$\max_{\tau\in[\tau_L,\tau_U]} \left\| \frac{d}{d\tau}\Sigma^{1/2}\boldsymbol{\beta}^*(\tau) \right\|_2 \le \underline{f}^{-1} \max_{\tau\in[\tau_L,\tau_U]} \|\mathbb{E}(\Sigma^{-1/2}\boldsymbol{x})\|_2 \le \underline{f}^{-1},$$

which, together with the mean value theorem, implies

$$||\boldsymbol{\beta}^*(\tau) - \boldsymbol{\beta}^*(\tau')||_\Sigma \le \underline{f}^{-1}|\tau - \tau'| \quad \text{for any} \quad \tau, \tau' \in [\tau_L, \tau_U]. \tag{4.10}$$

By the definitions in Condition 4.3.3, $G(u|\boldsymbol{x}) \le F(u|\boldsymbol{x})$ for any $u > 0$. Recall that we have assumed no censored observations at the low quantiles with $\tau \le \tau_L$. Hence, $G(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*(\tau_L)|\boldsymbol{x}) = F(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*(\tau_L)|\boldsymbol{x}) = \tau_L$, and $G(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*(\tau)|\boldsymbol{x}) \le \tau \le F(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*(\tau)|\boldsymbol{x})$ for $\tau_L < \tau \le \tau_U$. Condition 4.3.4 assures a fine grid by controlling the gap between two contiguous points, so that the approximation error does not exceed the statistical error.

4.3.2 Uniform rate of convergence and Bahadur representation

In this section, we characterize the statistical properties of the SEE estimators for censored quantile regression with growing dimensions. That is, the dimension $p = p_n$ is subject to the growth condition $p \asymp n^a$ for some $a \in (0,1)$. Our first result provides the uniform rate of convergence for the estimated coefficient function $\widehat{\boldsymbol{\beta}}(\cdot)$ under mild bandwidth constraints.

Theorem 4.3.1 (Uniform consistency)**.** *Assume Conditions 4.3.1–4.3.4 hold, and choose the bandwidth $h = h_n \asymp \{(p+\log n)/n\}^\gamma$ for some $\gamma \in [1/4, 1/2)$. Furthermore, let $n \gtrsim \{\zeta_p^2(p+$*

$\log n)^{1/2-\gamma}\}^{1/(1-\gamma)}$. *Then, the SEE estimator* $\widehat{\boldsymbol{\beta}}(\cdot) : [\tau_L, \tau_U] \mapsto \mathbb{R}^p$ *satisfies*

$$\sup_{\tau\in[\tau_L,\tau_U]} \|\widehat{\boldsymbol{\beta}}(\tau) - \boldsymbol{\beta}^*(\tau)\|_\Sigma \lesssim \left(\frac{1-\tau_L}{1-\tau_U}\right)^{C_0\overline{f}/\underline{g}} \underline{g}^{-1}\sqrt{\frac{p+\log n}{n}} \tag{4.11}$$

with probability at least $1 - C_1 n^{-1}$, *where* $C_0, C_1 > 0$ *are constants independent of* (n, p).

Since the deviation bound in (4.11) depends explicitly on n, p as well as other model parameters, this non-asymptotic result implies the classical asymptotic consistency by letting $n \to \infty$ with p fixed. From an asymptotic perspective, Theorem 4.3.1 implies that the smoothed estimator with a bandwidth $h = h_n \asymp \{\log(n)/n\}^\gamma$ for some $\gamma \in [1/4, 1/2)$ satisfies $\sup_{\tau_L\le\tau\le\tau_U} \|\widehat{\boldsymbol{\beta}}(\tau) - \boldsymbol{\beta}^*(\tau)\|_2 \to 0$ in probability as $n \to \infty$.

Recall that in the sequential estimation procedure described in Section 4.2.2, the j-th estimator $\widehat{\boldsymbol{\beta}}_j$ $(j \ge 1)$ depends implicitly on its predecessors through the estimating function (4.4). In other words, the accumulative estimation errors of $\widehat{\boldsymbol{\beta}}(\tau)$ for $\tau_L \le \tau < \tau_j$ may have a non-negligible impact on $\widehat{\boldsymbol{\beta}}_j = \widehat{\boldsymbol{\beta}}(\tau_j)$. The next result explicitly quantifies this accumulative error. For $\tau \in [\tau_L, \tau_U]$, define $p \times p$ matrices

$$\mathbf{J}(\tau) = \mathbb{E}\big\{g(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*(\tau)|\boldsymbol{x})\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\big\} \quad \text{and} \quad \mathbf{H}(\tau) = \mathbb{E}\big\{f(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*(\tau)|\boldsymbol{x})\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\big\}, \tag{4.12}$$

both of which are positive definite under Conditions 4.3.2 and 4.3.3. Moreover, define the integrated covariate effect and its estimate

$$\begin{aligned} \boldsymbol{\beta}^*_{\mathrm{int}}(\tau) &:= \mathbf{J}(\tau)\boldsymbol{\beta}^*(\tau) + \int_{\tau_L}^{\tau} \mathbf{H}(u)\boldsymbol{\beta}^*(u)\,\mathrm{d}H(u) \\ \text{and} \quad \widehat{\boldsymbol{\beta}}_{\mathrm{int}}(\tau) &:= \mathbf{J}(\tau)\widehat{\boldsymbol{\beta}}(\tau) + \int_{\tau_L}^{\tau} \mathbf{H}(u)\widehat{\boldsymbol{\beta}}(u)\,\mathrm{d}H(u), \end{aligned}$$

respectively, so that $\widehat{\boldsymbol{e}}(\tau) := \widehat{\boldsymbol{\beta}}_{\mathrm{int}}(\tau) - \boldsymbol{\beta}^*_{\mathrm{int}}(\tau)$ can be interpreted as the accumulated error in the

sequential estimation procedure up to τ. That is,

$$\widehat{\boldsymbol{e}}(\tau) = \underbrace{\mathbf{J}(\tau)\{\widehat{\boldsymbol{\beta}}(\tau) - \boldsymbol{\beta}^*(\tau)\}}_{\text{current step}} + \underbrace{\int_{\tau_L}^{\tau} \mathbf{H}(u)\{\widehat{\boldsymbol{\beta}}(u) - \boldsymbol{\beta}^*(u)\}\,\mathrm{d}H(u)}_{\text{preceding steps}}. \tag{4.13}$$

The following theorem provides a uniform Bahadur representation for $\widehat{\boldsymbol{e}}(\cdot)$.

Theorem 4.3.2 (Uniform Bahadur representation)**.** *Assume that the same set of conditions in Theorem 4.3.1 hold. Moreover, assume $\delta^* \asymp n^{-(1/2+\alpha)}$ for some $\alpha \in (0, 1/2)$. Then, the SEE estimator $\widehat{\boldsymbol{\beta}}(\cdot) : [\tau_L, \tau_U] \mapsto \mathbb{R}^p$ satisfies*

$$\widehat{\boldsymbol{e}}(\tau) = \widehat{\boldsymbol{\beta}}_{\text{int}}(\tau) - \boldsymbol{\beta}^*_{\text{int}}(\tau) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{U}_i(\tau) + \boldsymbol{r}_n(\tau), \tag{4.14}$$

where

$$\boldsymbol{U}_i(\tau) := \left\{ \tau_L + \int_{\tau_L}^{\tau} \bar{K}_h(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*(u))\,\mathrm{d}H(u) - \Delta_i \bar{K}_h(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*(\tau) - y_i) \right\} \boldsymbol{x}_i \tag{4.15}$$

satisfies $\sup_{\tau\in[\tau_L,\tau_U]} \|\mathbb{E}\boldsymbol{U}_i(\tau)\|_{\Sigma^{-1}} \lesssim h^2$, and the remainder process $\boldsymbol{r}_n(\cdot) : [\tau_L, \tau_U] \mapsto \mathbb{R}^p$ is such that

$$\sup_{\tau\in[\tau_L,\tau_U]} \|\boldsymbol{r}_n(\tau)\|_{\Sigma^{-1}} \lesssim m_4^{1/2}\frac{p+\log n}{nh^{1/2}} + m_3\frac{p+\log n}{n} + h\sqrt{\frac{p+\log n}{n}} + n^{-1/2-\alpha} \tag{4.16}$$

with probability at least $1 - C_2 n^{-1}$ for some absolute constant $C_2 > 0$, where m_q ($q = 3, 4$) are given in (4.9)*.*

Remark 4.3.1. *Together, the above uniform Bahadur representation and the production integration theory [Gill and Johansen, 1990] establish the asymptotic distribution of $\widehat{\boldsymbol{\beta}}(\cdot)$. Define*

$$\boldsymbol{\theta}^*(\tau) = \mathbf{J}(\tau)\boldsymbol{\beta}^*(\tau), \quad \widehat{\boldsymbol{\theta}}(\tau) = \mathbf{J}(\tau)\widehat{\boldsymbol{\beta}}(\tau) \quad \textit{and} \quad \boldsymbol{\Psi}(\tau) = \frac{1}{1-\tau}\mathbf{H}(\tau)\mathbf{J}(\tau)^{-1}, \quad \tau \in [\tau_L, \tau_U].$$

Then, equation (4.13) *reads* $\widehat{\boldsymbol{e}}(\tau) = \widehat{\boldsymbol{\theta}}(\tau) - \boldsymbol{\theta}^*(\tau) + \int_{\tau_L}^{\tau} \boldsymbol{\Psi}(u)\{\widehat{\boldsymbol{\theta}}(u) - \boldsymbol{\theta}^*(u)\}\mathrm{d}u$. *Combined with Theorem 4.3.2, this implies*

$$n^{1/2}\{\widehat{\boldsymbol{\theta}}(\tau) - \boldsymbol{\theta}^*(\tau)\} + \int_{\tau_L}^{\tau} \boldsymbol{\Psi}(u)\, n^{1/2}\{\widehat{\boldsymbol{\theta}}(u) - \boldsymbol{\theta}^*(u)\}\,\mathrm{d}u \\ = \frac{1}{n^{1/2}} \sum_{i=1}^{n} \{\boldsymbol{U}_i(\tau) - \mathbb{E}\boldsymbol{U}_i(\tau)\} + \bar{\boldsymbol{r}}_n(\tau), \quad \tau \in [\tau_L, \tau_U], \tag{4.17}$$

where the rescaled remainder $\bar{\boldsymbol{r}}_n(\cdot)$ *satisfies* $\sup_{\tau\in[\tau_L,\tau_U]} \|\bar{\boldsymbol{r}}_n(\tau)\|_2 = o_{\mathbb{P}}(1)$, *with a properly chosen bandwidth that will be discussed in Remark 4.3.2. Note that equation* (4.17) *is a stochastic differential equation for* $n^{1/2}\{\widehat{\boldsymbol{\theta}}(\tau) - \boldsymbol{\theta}^*(\tau)\}$ *[Peng and Huang, 2008]. From the classical production integration theory (Gill and Johansen [1990] and Section II.6 of Andersen et al. [1993]), it follows that*

$$n^{1/2}\{\widehat{\boldsymbol{\theta}}(\tau) - \boldsymbol{\theta}^*(\tau)\} = \boldsymbol{\phi}\left(\frac{1}{n^{1/2}} \sum_{i=1}^{n} \{\boldsymbol{U}_i(\tau) - \mathbb{E}\boldsymbol{U}_i(\tau)\}\right) + o_{\mathbb{P}}(1), \tag{4.18}$$

where $\boldsymbol{\phi}$ *is a linear operator from* $\mathscr{F}$ *to* $\mathscr{F}$ *defined as*

$$\boldsymbol{\phi}(\boldsymbol{g})(\tau) = \prod_{u\in[\tau_L,\tau]} \{\mathbf{I}_p - \boldsymbol{\Psi}(u)\mathrm{d}u\}\boldsymbol{g}(\tau_L) + \int_{\tau_L}^{\tau} \prod_{u\in(s,\tau]} \{\mathbf{I}_p - \boldsymbol{\Psi}(u)\mathrm{d}u\}\,\mathrm{d}\boldsymbol{g}(s) \tag{4.19}$$

for $\boldsymbol{g} \in \mathscr{F} := \{\boldsymbol{f} : [\tau_L, \tau_U] \to \mathbb{R}^p \,|\, \boldsymbol{f}$ *is left-continuous with right limit*$\}$, *and* Π *denotes the product integral; see Definition 1 in Gill and Johansen [1990]. After careful proofreading, we believe that the above form of* $\boldsymbol{\phi}(\cdot)$ *corrects an error (possibly a typo) in the proof of Theorem 2 in Peng and Huang [2008]; see the arguments between (B.1) and (B.3) therein. Specifically, the linear operator* $\boldsymbol{\phi}$ *in Peng and Huang [2008] reads*

$$\boldsymbol{\phi}(\boldsymbol{g})(\tau) = \prod_{u\in[\tau_L,\tau]} \{\mathbf{I}_p + \boldsymbol{\Psi}(u)\mathrm{d}u\}\boldsymbol{g}(\tau_L) + \int_{\tau_L}^{\tau} \prod_{u\in(s,\tau]} \{\mathbf{I}_p + \boldsymbol{\Psi}(u)\mathrm{d}u\}\,\mathrm{d}\boldsymbol{g}(s).$$

The asymptotic distribution of $n^{1/2}\{\widehat{\boldsymbol{\theta}}(\tau) - \boldsymbol{\theta}^*(\tau)\}$ *or its linear functional is thus deter-*

mined by that of

$$\boldsymbol{\phi}\left(\frac{1}{n^{1/2}}\sum_{i=1}^{n}\{\boldsymbol{U}_i(\tau)-\mathbb{E}\boldsymbol{U}_i(\tau)\}\right) \quad \textit{and} \quad \frac{1}{n^{1/2}}\sum_{i=1}^{n}\{\boldsymbol{U}_i(\tau)-\mathbb{E}\boldsymbol{U}_i(\tau)\}.$$

Remark 4.3.2 (Order of bandwidth)**.** *We further discuss the order of bandwidth h, as a function of* (n,p)*, required in Theorem 4.3.2 and Remark 4.3.1. Following* (4.17)*, if the moment parameters* m_3 *(absolute skewness) and* m_4 *(kurtosis) are dimension-free, the Bahadur linearization remainder* $\bar{\boldsymbol{r}}_n(\cdot)$ *satisfies with high probability that* $\sup_{\tau\in[\tau_L,\tau_U]}\|\bar{\boldsymbol{r}}_n(\tau)\|_{\Sigma^{-1}} \lesssim n^{1/2}h^2+(p+\log n)/(nh)^{1/2}+n^{-\alpha}$*. Set the bandwidth* $h \asymp \{(p+\log n)/n\}^{\gamma}$ *for some* $\gamma\in[1/4,1/2)$*, this implies*

$$\sup_{\tau\in[\tau_L,\tau_U]}\|\bar{\boldsymbol{r}}_n(\tau)\|_{\Sigma^{-1}} \lesssim \frac{(p+\log n)^{2\gamma}}{n^{2\gamma-1/2}}+\frac{(p+\log n)^{1-\gamma/2}}{n^{1/2-\gamma/2}}+\frac{1}{n^{\alpha}}=o_{\mathbb{P}}(1),$$

provided that $p=o(n^{1-1/(4\gamma)}\wedge n^{(1-\gamma)/(2-\gamma)})$*. In particular, letting* $1-1/(4\gamma)=(1-\gamma)/(2-\gamma)$ *yields* $\gamma=2/5$*. We therefore choose the bandwidth* $h\asymp\{(p+\log n)/n\}^{2/5}$*, so that all the asymptotic results (from uniform rate of convergence to Bahadur representation) hold under the growth condition* $p=o(n^{3/8})$ *of dimensionality p in sample size n.*

Theorem 4.3.2 explicitly characterizes the leading term of the integrated estimation error (4.13), along with a high probability bound on the remainder process. As discussed in Remark 4.3.1, the asymptotic distributions of $n^{1/2}\{\widehat{\boldsymbol{\beta}}(\tau)-\boldsymbol{\beta}^*(\tau)\}$ or its linear functional can be established based on the stochastic integral representation (4.18), which further depends on the centered random process $n^{-1/2}\sum_{i=1}^{n}\{\boldsymbol{U}_i(\cdot)-\mathbb{E}\boldsymbol{U}_i(\cdot)\}$. Let $\{\boldsymbol{a}_n\}_{n=1}^{\infty}$ be a sequence of deterministic vectors in $\mathbb{R}^p$, and define

$$\mathbb{G}_n(\tau):=\frac{1}{n^{1/2}}\sum_{i=1}^{n}\langle \boldsymbol{a}_n/||\boldsymbol{a}_n||_{\Sigma},\boldsymbol{U}_i(\tau)-\mathbb{E}\boldsymbol{U}_i(\tau)\rangle, \quad \tau\in[\tau_L,\tau_U]. \tag{4.20}$$

The asymptotic behavior of $\{\mathbb{G}_n(\tau):\tau\in[\tau_L,\tau_U]\}$ is provided in the following result.

Theorem 4.3.3 (Weak convergence)**.** *Assume Conditions 4.3.1–4.3.4 hold with* $\delta^* \asymp n^{-(1/2+\alpha)}$ *for some* $\alpha \in (0, 1/2)$*. Moreover, assume* $h \asymp \{(p+\log n)/n\}^{2/5}$ *and* $p = o(n^{3/8})$ *as* $n \to \infty$*. For any deterministic sequence of vectors* $\{\boldsymbol{a}_n\}_{n\geq 1}$*, if the following limit*

$$H(\tau, \tau') := \lim_{n\to\infty} \frac{1}{\|\boldsymbol{a}_n\|_{\Sigma}^2} \boldsymbol{a}_n^{\mathrm{T}} \mathbb{E}\big\{\boldsymbol{U}_i(\tau)\boldsymbol{U}_i(\tau')^{\mathrm{T}}\big\}\boldsymbol{a}_n \tag{4.21}$$

exists for any $\tau, \tau' \in [\tau_L, \tau_U]$ *with* $\boldsymbol{U}_i(\cdot)$ *defined in* (4.15)*, then*

$$\mathbb{G}_n(\cdot) \rightsquigarrow \mathbb{G}(\cdot) \quad \text{in } \ell^\infty([\tau_L, \tau_U]), \tag{4.22}$$

where $\mathbb{G}_n(\cdot)$ *is given in* (4.20)*, and* $\mathbb{G}(\cdot)$ *is a tight zero-mean Gaussian process with covariance function* $H(\cdot,\cdot)$ *and has almost surely continuous sample paths.*

Regarding the relative efficiency of the SEE estimator compared to its non-smoothed counterpart [Peng and Huang, 2008], note that the (integrated) kernel $\bar{K}_h(u)$ converges to $\mathbb{1}(u \geq 0)$ as $h \to 0$. Hence, the smoothed process $n^{-1/2}\sum_{i=1}^n \boldsymbol{U}_i(\tau)$ with $\boldsymbol{U}_i(\tau)$ given in (4.15) has the same asymptotic distribution as

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ \tau_L + \int_{\tau_L}^{\tau} \mathbb{1}(y_i \geq \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*(u))\, \mathrm{d}H(u) - \Delta_i \mathbb{1}(y_i \leq \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*(\tau)) \right\} \boldsymbol{x}_i.$$

As a result, the covariance function $H(\cdot,\cdot)$ defined in (4.21) coincides with that in Peng and Huang [2008]; see $\boldsymbol{\Sigma}(\cdot,\cdot)$ in the proof of Theorem 2 therein. In other words, the SEE estimator and Peng and Huang's estimator converge to the same Gaussian process as $n \to \infty$ with p fixed, and thence the asymptotic relative efficiency is 1. The technical devices required to deal with the fixed-p and growing-p cases are quite different. For the former, the consistency follows from the Glivenko-Cantelli theorem, and the weak convergence is a consequence of Donsker's theorem. To establish non-asymptotic results, we rely on a localized analysis as well as a (local) restricted strong convexity of the smoothed objective function that holds with high probability. The weak

convergence is based on the non-asymptotic uniform Bahadur representation (Theorem 4.3.2), complemented by showing the convergence of finite-dimensional marginals and the asymptotic tightness.

4.3.3 Rademacher multiplier bootstrap inference

In this section, we establish the theoretical guarantees of the Rademacher multiplier bootstrap for censored quantile regression as described in Section 4.2.3. In this case, $W_i = e_i + 1$ and e_i's are i.i.d. Rademacher random variables. For the random covaviate vector $\boldsymbol{x} \in \mathbb{R}^p$, we assume that the moment parameters m_3 and m_4 defined in (4.9) are dimension-free. We first present the (conditional) uniform consistency of the bootstrapped process $\{\widehat{\boldsymbol{\beta}}^\flat(\tau) : \tau \in [\tau_L, \tau_U]\}$ given the observed data $\mathbb{D}_n = \{(y_i, \Delta_i, \boldsymbol{x}_i\}_{i=1}^n$. Let $\mathbb{P}^*(\cdot) = \mathbb{P}(\cdot \,|\, \mathbb{D}_n)$ be the conditional probability given $\mathbb{D}_n$.

Theorem 4.3.4 (Conditional uniform consistency)**.** *Assume Conditions 4.3.1–4.3.4 hold, and let the bandwidth satisfy $h = h_n \asymp \{(p+\log n)/n\}^\gamma$ for some $\gamma \in [1/4, 1/2)$. Then, there exists an event $\mathscr{E} = \mathscr{E}(\mathbb{D}_n)$ with $\mathbb{P}(\mathscr{E}) \geq 1 - C_3 n^{-1}$ such that conditional on $\mathscr{E}$, the bound* (4.11) *holds, and the bootstrapped process $\widehat{\boldsymbol{\beta}}^\flat(\cdot) : [\tau_L, \tau_U] \mapsto \mathbb{R}^p$ satisfies*

$$\sup_{\tau \in [\tau_L, \tau_U]} ||\widehat{\boldsymbol{\beta}}^\flat(\tau) - \widehat{\boldsymbol{\beta}}(\tau)||_\Sigma \lesssim \sqrt{\frac{p+\log n}{n}}, \tag{4.23}$$

with $\mathbb{P}^*$-probability at least $1 - C_3 n^{-1}$, provided $\zeta_p^2 (p+\log n)^{1/2-\gamma} (p \log n)^{1/2} \lesssim n^{1-\gamma}$. Here $C_3 > 0$ is an absolute constant.

Analogously to (4.13), define the bootstrapped integrated error as

$$\widehat{\boldsymbol{e}}^\flat(\tau) := \mathbf{J}(\tau)\{\widehat{\boldsymbol{\beta}}^\flat(\tau) - \widehat{\boldsymbol{\beta}}(\tau)\} + \int_{\tau_L}^{\tau} \mathbf{H}(u)\{\widehat{\boldsymbol{\beta}}^\flat(u) - \widehat{\boldsymbol{\beta}}(u)\} \, \mathrm{d}H(u), \tag{4.24}$$

where $\mathbf{J}(\cdot)$ and $\mathbf{H}(\cdot)$ are given in (4.12). We then develop a linear representation for $\widehat{\boldsymbol{e}}^\flat(\tau)$, which can be viewed as a parallel version of Theorem 4.3.2 in the bootstrap world.

Theorem 4.3.5 (Conditional uniform Bahadur representation)**.** *Assume the conditions in Theorem 4.3.4 hold, and that the kernel $K(\cdot)$ in Condition 4.3.1 is Lipschitz continuous. Moreover, assume $\delta^* \lesssim n^{-(1/2+\alpha)}$ for some $\alpha > 0$. Then, there exists an event $\mathscr{F} = \mathscr{F}(\mathbb{D}_n)$ with $\mathbb{P}(\mathscr{F}) \geq 1 - C_4 n^{-1}$ such that conditional on $\mathscr{F}$,* (4.14)–(4.16) *hold, and the bootstrapped process $\widehat{\boldsymbol{\beta}}^\flat(\cdot) : [\tau_L, \tau_U] \mapsto \mathbb{R}^p$ satisfies*

$$\widehat{\boldsymbol{e}}^\flat(\tau) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{U}_i^\flat(\tau) + \boldsymbol{r}_n^\flat(\tau), \tag{4.25}$$

where $\boldsymbol{U}_i^\flat(\tau) = e_i \boldsymbol{U}_i(\tau)$ with $\boldsymbol{U}_i(\tau)$ defined in (4.15)*, and*

$$\begin{aligned} &\sup_{\tau\in[\tau_L,\tau_U]} \|\boldsymbol{r}_n^\flat(\tau)\|_{\Sigma^{-1}} \\ &\lesssim m_4^{1/2}\frac{p+\log n}{nh^{1/2}} + h\sqrt{\frac{p+\log n}{n}} + \zeta_p^2\frac{(p+\log n)(p\log n)^{1/2}}{n^{3/2}h} + n^{-1/2-\alpha} \end{aligned} \tag{4.26}$$

with $\mathbb{P}^*$-probability at least $1 - C_4 n^{-1}$.

Theorem 4.3.5 shows that the bootstrap integrated error $\widehat{\boldsymbol{e}}^\flat(\cdot)$ can be approximated, up to a higher order remainder, by the linear process $\{(1/n)\sum_{i=1}^n e_i \boldsymbol{U}_i(\tau) : \tau \in [\tau_L, \tau_U]\}$, where e_i's are independent Rademacher random variables, and $\mathbb{E}^*\boldsymbol{U}_i^\flat(\tau) = \boldsymbol{0}$. Provided that $h \asymp \{(p+\log n)/n\}^{2/5}$ and p satisfies the growth condition $p = o(n^{3/8})$ as in Theorem 4.3.3, then applying the same analysis in Remark 4.3.1 gives us the following stochastic integral representation: with probability (over $\mathbb{D}_n$) approaching one, $\sup_{\tau\in[\tau_L,\tau_U]} \|\boldsymbol{r}_n^\flat(\tau)\|_{\Sigma^{-1}} = o_{\mathbb{P}^*}(1)$, and

$$n^{1/2}\mathbf{J}(\tau)\{\widehat{\boldsymbol{\beta}}^\flat(\tau) - \widehat{\boldsymbol{\beta}}(\tau)\} = \boldsymbol{\phi}\left(\frac{1}{n^{1/2}}\sum_{i=1}^{n}\boldsymbol{U}_i^\flat(\tau)\right) + o_{\mathbb{P}^*}(1), \tag{4.27}$$

where $\boldsymbol{\phi}$ is the linear operator defined in (4.19). Note that $\mathbb{E}^*\{\boldsymbol{U}_i^\flat(s)\boldsymbol{U}_i^\flat(t)^{\mathrm{T}}\} = \boldsymbol{U}_i(s)\boldsymbol{U}_i(t)^{\mathrm{T}}$ for any $s,t \in [\tau_L, \tau_U]$. It can be shown that on $[\tau_L, \tau_U]$, $n^{-1/2}\sum_{i=1}^n\{\boldsymbol{U}_i(\cdot) - \mathbb{E}\boldsymbol{U}_i(\cdot)\}$ has the same asymptotic distribution as $n^{-1/2}\sum_{i=1}^n \boldsymbol{U}_i^\flat(\cdot)$ conditionally on the data $\mathbb{D}_n$; see Theorem 4.3.3 and

Theorem 4.3.6 below. This, together with (4.18) and (4.27), validates to some level the use of the bootstrap process $\widehat{\boldsymbol{\beta}}^{\flat}(\cdot)$ in the inference. To illustrate this, consider the following bootstrap counterpart of the process $\mathbb{G}_n(\cdot)$ defined in (4.20):

$$\mathbb{G}_n^{\flat}(\tau) := \frac{1}{n^{1/2}} \sum_{i=1}^{n} \langle \boldsymbol{a}_n / ||\boldsymbol{a}_n||_{\Sigma}, \boldsymbol{U}_i^{\flat}(\tau) \rangle, \quad \tau \in [\tau_L, \tau_U]. \tag{4.28}$$

Theorem 4.3.6 (Validation of bootstrap process)**.** *Assume Conditions 4.3.1–4.3.4 hold with* $\delta^* \lesssim n^{-(1/2+\alpha)}$ *for* $\alpha \in (0, 1/2)$, $h \asymp \{(p+\log n)/n\}^{2/5}$ *and* $p = o(n^{3/8})$. *In addition, assume the kernel* $K(\cdot)$ *is Lipschitz continuous. Then, for any sequence of (deterministic) vectors* $\{\boldsymbol{a}_n\}_{n=1}^{\infty}$, *there exists a sequence of events* $\{\mathscr{F}_n = \mathscr{F}_n(\mathbb{D}_n)\}_{n=1}^{\infty}$ *such that* $\mathbb{P}(\mathscr{F}_n) \to 1$, *and conditional on* $\{\mathscr{F}_n\}_{n=1}^{\infty}$, (4.25) *holds and the conditional distribution of* $\mathbb{G}_n^{\flat}(\cdot)$ *given* $\mathbb{D}_n$ *is asymptotically equivalent to the unconditional distribution of* $\mathbb{G}_n(\cdot)$ *established in* (4.22).

4.4 Regularized Censored Quantile Regression

We extend the proposed SEE approach to high-dimensional sparse QR models with random censoring. The goal is to identify the set of relevant predictors, defined as

$$\mathscr{S}^* = \bigcup_{\tau \in [\tau_L, \tau_U]} \operatorname{supp}\big(\boldsymbol{\beta}^*(\tau)\big), \tag{4.29}$$

assuming that its cardinality $s := |\mathscr{S}^*|$ is much smaller than the ambient dimension p—the total number of predictors, but may grow with sample size n. Recall the sequentially defined smoothed loss functions $\widehat{L}_k(\cdot)$ $(k = 0, 1, \ldots, m)$ in (4.5) and (4.6). When $p < n$, finding the solution to the SEE $\widehat{Q}_k(\boldsymbol{\beta}) = 0$ is equivalent to solving the optimization problem $\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \widehat{L}_k(\boldsymbol{\beta})$. For fitting sparse models in high dimensions, we start with the ℓ_1-penalized approach [Tibshirani, 1996, Belloni and Chernozhukov, 2011]. At quantile levels $\tau_L = \tau_0 < \tau_1 < \cdots < \tau_m = \tau_U$, we define

ℓ_1-penalized smoothed CQR estimators $\widehat{\boldsymbol{\beta}}_k := \widehat{\boldsymbol{\beta}}(\tau_k)$ sequentially as

$$\widehat{\boldsymbol{\beta}}(\tau_k) \in \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} \big\{ \widehat{L}_k(\boldsymbol{\beta}) + \lambda_k \cdot \|\boldsymbol{\beta}\|_1 \big\}, \tag{4.30}$$

for $k \in \{0,\ldots,m\}$, and define $\widehat{\boldsymbol{\beta}}(\tau) = \widehat{\boldsymbol{\beta}}(\tau_{k-1})$ for $\tau \in (\tau_{k-1},\tau_k)$. It is worth noticing that for each $k \geq 1$, $\widehat{L}_k(\cdot)$ is essentially a shifted or perturbed version of $\widehat{L}_0(\cdot)$, that is, $\widehat{L}_k(\boldsymbol{\beta}) = \widehat{L}_0(\boldsymbol{\beta}) - (1/n)\sum_{i=1}^n \sum_{j=0}^{k-1} \bar{K}_h(y_i - \boldsymbol{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_j)\{H(\tau_{j+1}) - H(\tau_j)\}\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}$, where $H(u) = -\log(1-u)$. All these empirical loss functions are convex, and have the same second-order properties on the Hessian.

Condition 4.4.1 (Random design in high dimensions)**.** *The covariate vector $\boldsymbol{x} = (x_1,\ldots,x_p)^{\mathrm{T}} \in \mathscr{X} \subseteq \mathbb{R}^p$ ($x_1 \equiv 1$) is compactly supported with $\max_{1\leq j\leq p}|x_j| \leq C_0$ almost surely for some $C_0 \geq 1$. For convenience, assume $C_0 = 1$. The normalized vector $\Sigma^{-1/2}\boldsymbol{x}$ has uniformly bounded kurtosis, that is, m_4 defined in* (4.9) *is a dimension-free constant, where $\Sigma = \mathbb{E}(\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}})$ is positive definite.*

Theorem 4.4.1. *Assume Conditions 4.3.1, 4.3.3, 4.3.4 and Condition 4.4.1 hold. Under the sample size scaling $n \gtrsim s^3 \log p$, let the bandwidth h and penalty levels λ_k's satisfy $s\sqrt{\log(p)/n} \lesssim h \lesssim \{s\log(p)/n\}^{1/4}$ and $\lambda_k \asymp \{1+\log(\frac{1-\tau_L}{1-\tau_k})\}\sqrt{\log(p)/n}$ for $k = 0,1,\ldots,m$. Then, there exist constants $C_1, C_2 > 0$ independent of (s,p,n) such that*

$$\sup_{\tau_L\leq\tau\leq\tau_U} \|\widehat{\boldsymbol{\beta}}(\tau) - \boldsymbol{\beta}^*(\tau)\|_\Sigma \leq C_1 \frac{\sigma}{\underline{g}\sqrt{\gamma_l}} \log\big(\tfrac{1-\tau_L}{1-\tau_U}\big)\sqrt{\frac{s\log p}{n}}$$

with probability at least $1 - C_2 p^{-1}$, where $\gamma_l = \lambda_{\min}(\Sigma)$, the minimal eigenvalue of Σ, and $\sigma = \max_{1\leq j\leq p}\sigma_{jj}$.

Theorem 4.4.1 provides the rate of convergence for the ℓ_1-penalized smoothed CQR estimator $\widehat{\boldsymbol{\beta}}(\cdot)$ uniformly in the set of quantile indices $\tau \in [\tau_L, \tau_U]$. Under a similar set of assumptions, Zheng, Peng and He [2018] established the uniform convergence rate for the ℓ_1-penalized (non-smoothed) CQR estimator, which is of order $\exp(Cs)\sqrt{s\log(p\vee n)/n}$. We conjecture that

the additional exponential term $\exp(Cs)$ is a consequence of the marginal smoothness condition posed in Zheng, Peng and He [2018] (see Condition (C4) therein), and can be relaxed as in our Theorem 4.4.1. In fact, our analysis relies on the global Lipschitz property (4.10), which follows directly from the model assumption (4.1) and a lower bound on the conditional density.

Remark 4.4.1 (Comments on the tuning parameters h and $\{\lambda_k\}_{k=0}^m$)**.** *To achieve the same convergence rate $\sqrt{s\log(p)/n}$ as for the ℓ_1-penalized QR estimator with non-censored data [Belloni and Chernozhukov, 2011], the bandwidth h is required to be in the range specified in Theorem 4.4.1; for example, one may choose $h \asymp \{s\log(p)/n\}^{1/4}$. Since such a choice depends on the unknown sparsity, in practice we simply choose h to be of order $\{\log(p)/n\}^{1/4}$. Since the numerical performance is rather insensitive to the choice of bandwidth, we suggest a default value $h = \max\{0.05, 0.5\{\log(p)/n\}^{1/4}\}$ in high dimensions although it can also be tuned by cross-validation.*

The penalty levels λ_k's play a more pivotal role in obtaining a reasonable fit for the whole CQR process. Our theoretical analysis suggests that $\{\lambda_k\}_{k=0}^m$ should be chosen as a slowly growing sequence along the τ-grid. Numerical results also confirm that a single λ value, even after proper tuning, cannot guarantee a quality estimation of the entire regression process. On the other hand, it is computationally prohibitive to determine each λ_k ($k = 0, 1, \ldots, m$) via cross-validation. By examining the proof of Theorem 4.4.1, we see that once λ_0 is specified, the subsequent λ_k's satisfy $\lambda_k = \{1 + \log(\frac{1-\tau_L}{1-\tau_k})\}\lambda_0$ for $k = 1, \ldots, m$. Therefore, to implement the proposed sequential procedure, we only treat λ_0 as a tuning parameter, and use the above formula to determine the rest of λ_k's.

Remark 4.4.2 (Adaptive ℓ_1-penalization)**.** *It has been recognized that the ℓ_1-penalized estimator, with the penalty level determined via cross-validation, typically has small prediction error but has a non-negligible estimation bias and tends to overfit with many false discoveries. To reduce the estimation error and false positives, a popular strategy is to use reweighted ℓ_1-penalization via either adaptive Lasso [Zou, 2006] or the local linear approximation (LLA) method for*

folded-concave penalties [Fan and Li, 2001, Zou and Li, 2008]. Let $w(\cdot)$ be a non-increasing and non-negative function defined on $[0,\infty)$. Fix k, let $\widehat{\boldsymbol{\beta}}_k^{(0)} = \widehat{\boldsymbol{\beta}}(\tau_k)$ be the ℓ_1-penalized censored QR estimator at quantile level τ_k. For $t = 1,\ldots,T$, we iteratively update the previous estimate $\widehat{\boldsymbol{\beta}}_k^{(t-1)}$ by solving

$$\widehat{\boldsymbol{\beta}}_k^{(t)} = (\widehat{\beta}_{k,1}^{(t)},\ldots,\widehat{\beta}_{k,p}^{(t)})^{\mathrm{T}} \in \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \left\{ \widehat{L}_k(\boldsymbol{\beta}) + \lambda_k \cdot \sum_{j=1}^{p} w\big(|\widehat{\beta}_{k,j}^{(t-1)}|/\lambda_k\big)|\beta_j| \right\}.$$

When $T = 1$ and $w(u) = u^{-1}$ for $u > 0$ (or $(u+\varepsilon)^{-1}$ for a small constant $\varepsilon > 0$), this corresponds to an adaptive Lasso-type estimator [Zou, 2006]; when $w(u) = \mathbb{1}(u \le 1) + \frac{(a-u)_+}{a-1}\mathbb{1}(u > 1)$ for $u \ge 0$ and some $a > 2$, this corresponds to the LLA method using the smoothly clipped absolute deviation (SCAD) penalty [Fan and Li, 2001]; when $w(u) = (1-u/a)_+$ for $u \ge 0$ and some $a \ge 1$, this corresponds to the LLA method using the minimax concave penalty (MCP) [Zhang, 2010].

4.5 Numerical Studies

We apply the proposed methods in Sections 4.2 and 4.4 on simulated datasets and compare to that of Peng and Huang [2008] and Zheng, Peng and He [2018] for both low- and high-dimensional settings in Sections 4.5.1 and 4.5.2, respectively. The proposed method involves selecting a smoothing parameter h: for $p < n$, we set $h = \{(p+\log n)/n\}^{2/5} \vee 0.05$; for $p > n$, guided by Remark 4.4.1, we set $h = \{0.05 \vee 0.5\{\log(p)/n\}^{1/4}\}$. We found that the performance of our proposed method is insensitive to the choice of bandwidth, as also observed in Fernandes, Guerre and Horta [2021] and He et al. [2022]. We implemented Peng and Huang [2008] using the `crq` function with `method = "PengHuang"` from the `quantreg` package [Koenker, 2008]. On the other hand, Zheng, Peng and He [2018] is implemented using the barebones function `LASSO.fit` from `rqPen` [Sherwood and Maidman, 2022] instead of the function `rq(..., method = "lasso")` in the package `quantreg`. This is because the

function `rq(..., method = "lasso")` reports some numerical issues (e.g., singular design error) frequently in our numerical studies. All of the numerical studies are performed on a worker node with 32 CPUs, 2.5 GHz processor, and 512 GB of memory in a high-performance computing cluster.

4.5.1 Censored quantile regression: estimation and inference

We assess the performance of our proposed method in the low-dimensional setting with $n = 5000$ and $p = 100$. We start with generating the random covariates $\widetilde{\boldsymbol{x}}_i \in \mathbb{R}^p$ from a mixture of different distributions to represent different types of variables commonly encountered in many datasets. In particular, we generate the first 45 covariates from $\mathcal{N}(\mathbf{0}, \Sigma = (\sigma_{jk})_{1 \le j,k \le 45})$, where $\sigma_{jk} = 0.5^{|j-k|}$ for $1 \le j,k \le 45$, the second 45 covariates from a multivariate uniform distribution on the cube $[-2,2]^{45}$ with the same covariance matrix Σ using the `R` package `MultiRNG`, and the last 10 covariates from a Bernoulli distribution. Note that the three blocks of covariates generated are independent across the blocks. The response variables $z_i \in \mathbb{R}$ are then generated from the following models, both of which satisfy the global assumption in (4.1).

(i) Homoscedastic model: $z_i = \langle \widetilde{\boldsymbol{x}}_i, \boldsymbol{\gamma} \rangle + \varepsilon_i$ for $i = 1, \ldots, n$, where $\gamma_j \sim \text{Uniform}(-2,2)$ for $j = 1, \ldots, p$. Let $Q_{t_2}(\tau)$ be the τ-quantile of the t_2-distribution, and let $\boldsymbol{x}_i = (1, \widetilde{\boldsymbol{x}}_i^{\mathrm{T}})^{\mathrm{T}}$. Then, the above model can be equivalently formulated as

$$z_i = \langle \boldsymbol{x}_i, \boldsymbol{\beta}^*(\tau) \rangle, \;\; i = 1, \ldots, n, \;\; \text{where} \;\; \boldsymbol{\beta}^*(\tau) = (Q_{t_2}(\tau), \boldsymbol{\gamma}^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{p+1}. \tag{4.31}$$

Under the above model, the covariate effects remain the same across all quantile levels.

(ii) Heteroscedastic model: $z_i = \langle \widetilde{\boldsymbol{x}}_i, \boldsymbol{\gamma} \rangle + |\widetilde{x}_{i,1}| \cdot \varepsilon_i$ for $i = 1, \ldots, n$, where $\gamma_1 = 0$ and $\gamma_j \sim \text{Uniform}(-2,2)$ for $j = 2, \ldots, p$. Let $\boldsymbol{x}_i = (1, |\widetilde{x}_{i,1}|, \widetilde{\boldsymbol{x}}_{i,-1}^{\mathrm{T}})^{\mathrm{T}}$, where $\widetilde{\boldsymbol{x}}_{i,-1} \in \mathbb{R}^{p-1}$ is obtained

by removing the first element of $\widetilde{\boldsymbol{x}}_i$. The model is equivalent to

$$z_i = \langle \boldsymbol{x}_i, \boldsymbol{\beta}^*(\tau) \rangle, \;\; i = 1, \ldots, n, \;\; \text{where} \;\; \boldsymbol{\beta}^*(\tau) = (0, Q_{t_2}(\tau), \gamma_2, \ldots, \gamma_p)^{\mathrm{T}} \in \mathbb{R}^{p+1}. \quad (4.32)$$

In this model, the first covariate has varying marginal effects for different quantile levels. Specifically, the effect of $|\widetilde{x}_1|$ on the τ-th quantile of z is $F_{t_2}^{-1}(\tau)$, which is negligible when $\tau \approx 0.5$, but grows stronger as τ moves towards 0 or 1.

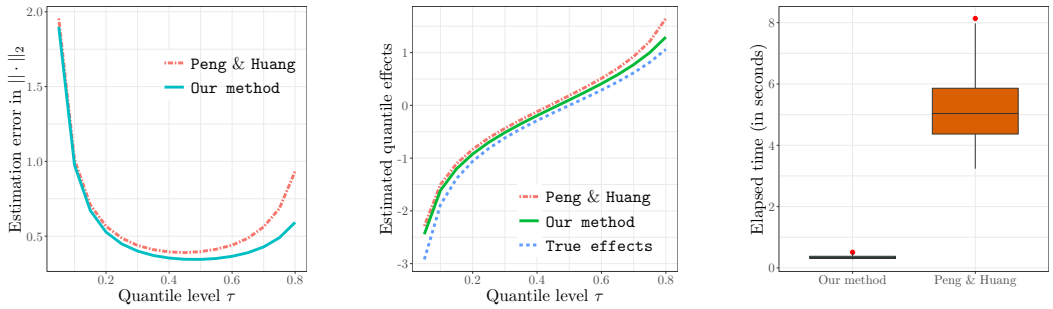
For both types of models, the random censoring variables are generated from a Gaussian mixture distribution, that is,

$$C_i \sim \mathbb{1}\{w_i = 1\}\mathcal{N}(0, 16) + \mathbb{1}\{w_i = 2\}\mathcal{N}(5, 1) + \mathbb{1}\{w_i = 3\}\mathcal{N}(10, 0.25) \quad (4.33)$$

for $i = 1, \ldots, n$, where w_i is sampled from $\{1, 2, 3\}$ with equal probability, and $y_i = z_i \wedge C_i$ is the censored outcome. The corresponding censoring rate varies from 25% to 50%.

We implement both methods with a quantile grid of $\{\tau_k\}_{k=0}^m = \{0.05, 0.1, \ldots, 0.75, 0.8\}$. At each quantile level τ_k, we use the estimation error under the ℓ_2 norm, $\|\widehat{\boldsymbol{\beta}}(\tau_k) - \boldsymbol{\beta}^*(\tau_k)\|_2$, as a general measure of accuracy. We also calculate the run-time in seconds for both methods. Results, averaged across 500 independent replications, are reported in Figure 4.1. Figures 4.1(a) and (d) contain the estimation error under the ℓ_2 norm across all quantile levels; Figures 4.1(b) and (d) contains the regression coefficient that varies across quantile levels, i.e., $\{\beta_0(\tau_k)\}_{k=0}^m$ for model (4.31) and $\{\beta_1(\tau_k)\}_{k=0}^m$ for model (4.32); and Figures 4.1(c) and (f) contain the computation time for fitting the entire QR process. We see that the two methods perform very closely at low quantile levels, and the smoothed approach is particularly advantageous at high quantile levels. Computationally, our implementation of the smoothed method is about 10 to 20 times faster than Peng and Huang [2008]'s method, implemented by the `crq` function in `quantreg`.

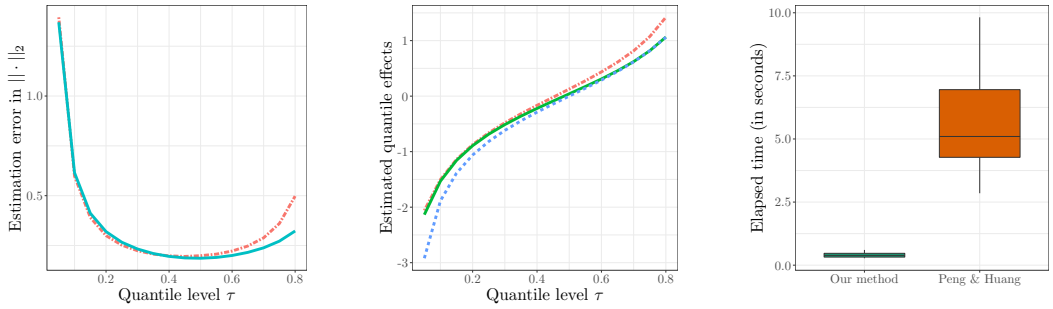
Next, we consider both the proposed multiplier bootstrap detailed in Section 4.2.3 and the classical paired bootstrap for performing statistical inference at $\tau = 0.5$. Three types of 95%



(a) ℓ_2-error under model (4.31)

(b) Estimated quantile effects under model (4.31)

(c) Runtime under model (4.31)



(d) ℓ_2-error under model (4.32)

(e) Estimated quantile effects under model (4.32)

(f) Runtime under model (4.32)

Figure 4.1. Numerical comparisons among CQR and our smoothed CQR for models (4.31)–(4.32) along the quantile grid. The left panels (a) and (d) display the ℓ_2-induced estimation errors $\|\widehat{\boldsymbol{\beta}}(\tau_k) - \boldsymbol{\beta}^*(\tau_k)\|_2$. The middle panels (b) and (e) present the estimated quantile effects, which are $\widehat{\beta}_0(\tau_k)$ in model (4.31) and $\widehat{\beta}_1(\tau_k)$ in model (4.32) accordingly. The blue dashed lines in the middle panels represent the true quantile effects $Q_{t_2}(\tau)$. The right panels (c) and (f) record the empirical running time of the processes along the grid points.

confidence intervals (CIs) are constructed with $B = 1000$ bootstrap samples: the percentile CI, the pivotal CI, and the normal CI. Coverage proportions for all of the covariates, confidence interval width for the first covariate, and computational time for the entire bootstrap process, averaged over 500 replications, are plotted in Figure 4.2. Under the homogeneous setting (4.31), all types of confidence intervals produced by multiplier bootstrap maintain the nominal level, while the normal intervals by pair resampling suffer from under coverage. In the heterogeneous setting (4.32), although outliers that correspond to the confidence intervals for the first covariate exist for both methods, multiplier bootstrap manages to mitigate this issue. Furthermore, compared to pair resampling, multiplier bootstrap constructs narrower confidence intervals with slightly

smaller standard deviations. Finally, the computational advantage of multiplier bootstrap for smoothed CQR is evident in Figures 4.2(c) and (f).

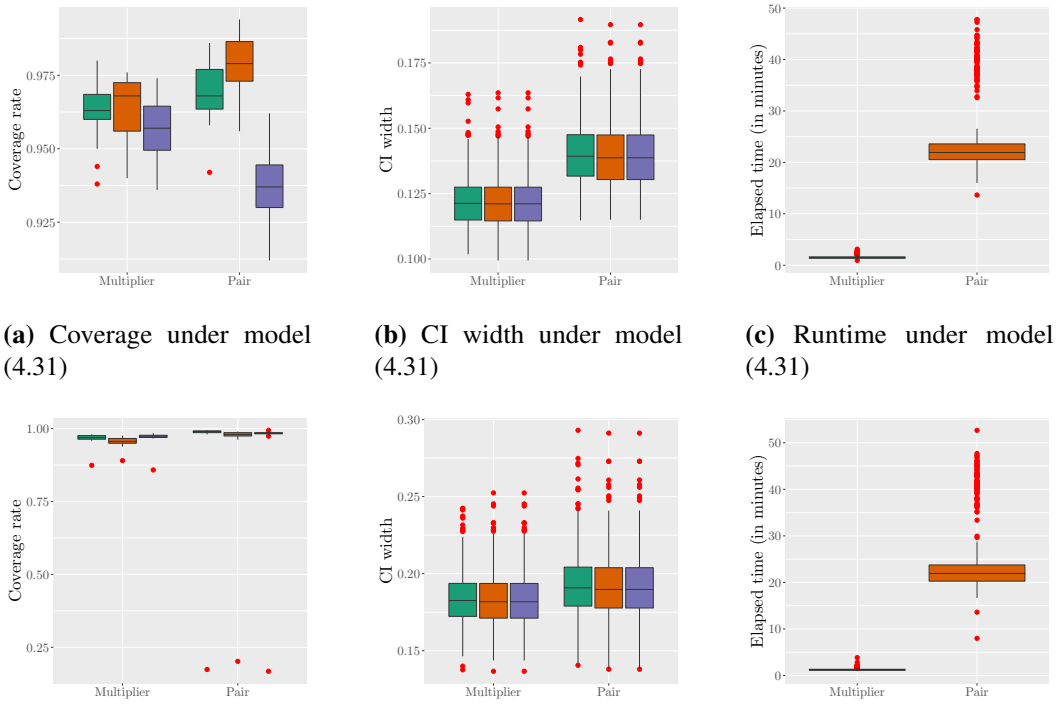
To better appreciate the computational advantage of smoothed CQR, we further consider large-scale simulation settings by setting $n \in \{1000, 2000, \ldots, 20000\}$ and $p = n/100$. We use the same data generating processes as in (4.31)–(4.33), except that the covariates $\widetilde{\boldsymbol{x}}_i$ are now generated from $\mathcal{N}(\mathbf{0}_p, \Sigma)$ with $\Sigma = (0.5^{|j-k|})_{1\le j,k\le p}$. The censoring rate varies from 30% to 45%. In this case, we restrict attention to the estimation error and runtime of the two methods when $\tau = 0.7$. The results, averaged over 500 repetitions, are presented in Figure 4.3. We see from Figure 4.3 that the computation gain of the proposed method over Peng and Huang [2008] is dramatic, without compromising the statistical accuracy.

4.5.2 High-dimensional censored quantile regression

In this section, we examine the numerical performance of the regularized smoothed CQR method with different penalties, which will also be compared with its non-smoothed counterpart [Zheng, Peng and He, 2018]. For the smoothed method, we consider both the ℓ_1 and folded-concave penalties (SCAD and MCP). The latter is implemented by the LLA algorithm as described in Remark 4.4.2. The computational details are described in Section A.2 of the supplementary material.

Penalized CQR involves selecting a sequence of regularization parameters $\{\lambda_k\}_{k=0}^m$ that correspond to the predetermined τ-grid $\{\tau_k\}_{k=0}^m$. Guided by Theorem 4.4.1 and Remark 4.4.1, we adopt a sequence of dilating λ_k's with $\lambda_k = \{1 + \log(\frac{1-\tau_L}{1-\tau_k})\}\lambda_0$ for $k = 1, \ldots, m$, where λ_0 is chosen via the K-fold cross-validation ($K = 3$ in our studies). To accommodate censoring, the cross-validation criterion is based on the the empirical mean of deviance residuals [Therneau, Grambsch and Fleming, 1990]

$$R(\lambda) := \frac{1}{n}\frac{1}{m+1}\sum_{i=1}^{n}\sum_{k=0}^{m}\sqrt{-2\{M_i(\tau_k,\lambda) + \Delta_i \log(\Delta_i - M_i(\tau_k,\lambda))\}} \tag{4.34}$$



(a) Coverage under model (4.31)

(b) CI width under model (4.31)

(c) Runtime under model (4.31)

(d) Coverage under model (4.32)

(e) CI width under model (4.32)

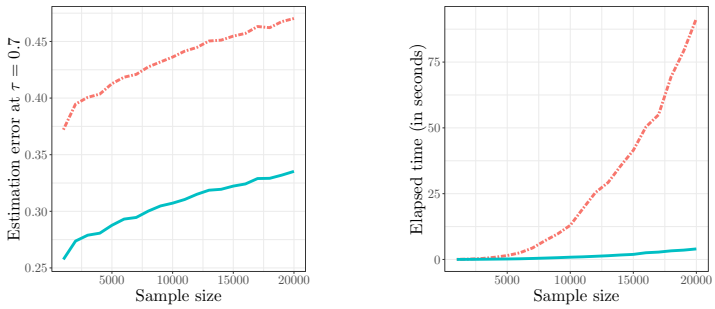
(f) Runtime under model (4.32)

Figure 4.2. Box plots of the empirical coverage, confidence interval width, and running time for two resampling-based methods. "Multiplier" refers to the proposed multiplier bootstrap method, and "Pair" refers to pair resampling with replacement in the regression setting. In panels (a), (b), (d), and (e), within each method, different colors of boxes represent different types of confidence interval: (i) green boxes for percentile interval, (ii) orange boxes for pivotal interval, and (iii) purple boxes for normal interval.

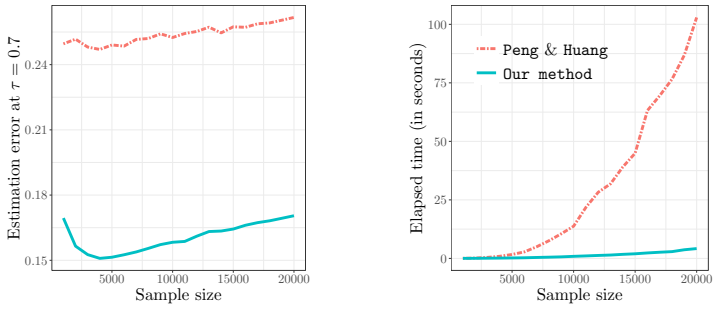
on the validation set, where

$$M_i(\tau_k,\lambda) = \mathbb{1}\{y_i \le \boldsymbol{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}(\tau_k,\lambda), \Delta_i = 1\} - \int_{\tau_0}^{\tau_k} \mathbb{1}\{y_i \ge \boldsymbol{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}(u,\lambda)\}\,\mathrm{d}H(u) - \tau_0$$

for $k = 0,\ldots,m$ are the martingale residuals and $\widehat{\boldsymbol{\beta}}(\tau,\lambda)$ refers to the estimated $\boldsymbol{\beta}(\tau)$ with a dilating λ_k's starting with $\lambda_0 = \lambda$. The deviance (4.34) produces a more symmetric distribution through a transformation on the skewed martingale residuals, and is also used in Zheng, Peng and He [2018] and Fei et al. [2021]. In our simulations, we choose λ_0 from 50 candidates equally spaced on the interval $[0.01, 0.2]$.



(a) ℓ_2-error under model (4.31) **(b)** Runtime under model (4.31)



(c) ℓ_2-error under model (4.32) **(d)** Runtime under model (4.32)

Figure 4.3. Numerical comparisons between CQR and smoothed CQR under models (4.31) and (4.32) with increasing (n, p) subject to $p = n/100$. The left panels (a) and (c) display the ℓ_2-error at $\tau = 0.7$ versus sample size. The right panels (b) and (d) present the runtime (in second) versus sample size.

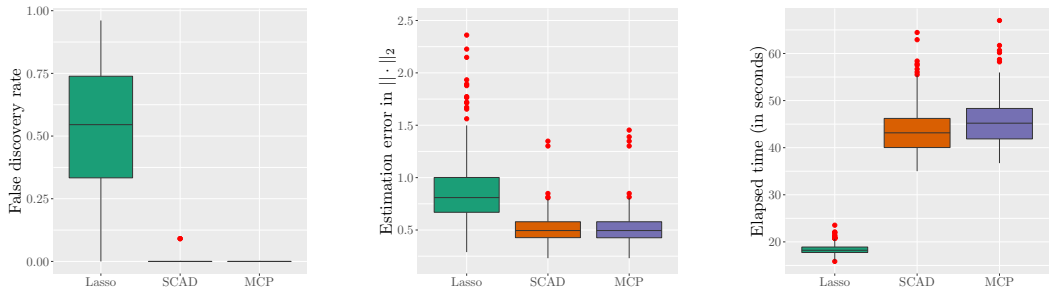
In all of our numerical studies, we generate covariates $\widetilde{\boldsymbol{x}}_i \in \mathbb{R}^p$ from $\mathcal{N}(\mathbf{0}, \Sigma)$, where Σ is as defined in Section 4.5.1, and the random errors $\varepsilon_i \sim t_2$. The response variables z_i are generated from Models (4.31)–(4.32), but with different $\boldsymbol{\gamma}$. For Model (4.31), we consider a sparse $\boldsymbol{\gamma}$ with global sparsity $s = 10$ by setting $\gamma_j \sim \text{Uniform}(1, 1.5)$ for $j = 1, \ldots, 10$, and the rest to be zero. For Model (4.32), $\boldsymbol{\gamma}$ is generated similarly except with $\gamma_1 = 0$. The random censoring variables are generated from (4.33), with overall censoring rates approximately 25%–30%.

Since the estimated active set depends on the entire quantile process, all numerical experiments are conducted via an estimation-after-selection procedure [Zheng, Peng and He, 2018]. That is, in stage one, we perform regularized smoothed CQR to obtain the set $\widehat{\mathscr{S}} = \cup_{\tau \in \{\tau_0, \ldots, \tau_m\}} \text{supp}\big(\widehat{\boldsymbol{\beta}}(\tau)\big)$. In stage two, we perform smoothed CQR using the covariates in $\widehat{\mathscr{S}}$. Recall that $\mathscr{S}$ is the true active set defined in (4.29), and let $\mathscr{S}^{\text{c}}$ be its complement. To assess

the numerical performance of our proposed method, we report (1) the true positive rate (TPR), $\text{TPR} = |\mathscr{S} \cap \widehat{\mathscr{S}}| \,/\, |\mathscr{S}|$; (2) the false discovery rate (FDR), $\text{FDR} = |\mathscr{S}^{\text{c}} \cap \widehat{\mathscr{S}}| \,/\, |\widehat{\mathscr{S}}|$; (3) average ℓ_2-error, $(1/m)\sum_{k=0}^{m} \|\widehat{\boldsymbol{\beta}}(\tau_k) - \boldsymbol{\beta}(\tau_k)\|_2$; and (4) elapsed time for running the estimation-after-selection process, including cross-validation.

Results for the proposed method using different penalty functions, averaged over 500 replications when $(n, p) = (400, 1000)$, are reported in Figure 4.4. As expected, ℓ_1-penalized method tends to select larger models with many spurious variables, and thus has higher false discovery rates than SCAD and MCP. Under the heterogeneous model, both SCAD and MCP sometimes miss the first true signal and have lower TPR than Lasso. This is due to the fact that the first signal corresponds to the evolving quantile effect $Q_{t_2}(\tau)$ that vanishes as τ approaches 0.5, and therefore is more likely to be missed by folded-concave regularization.

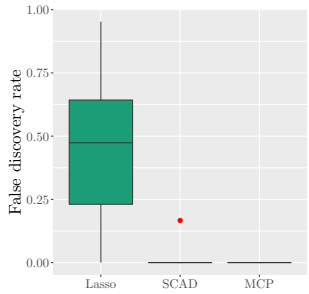
To better demonstrate the computational efficiency of our method on large-scale data, we consider the ℓ_1-penalized CQR (CQR-Lasso) method [Zheng, Peng and He, 2018] as a benchmark. As discussed in Zheng, Peng and He [2018], CQR-Lasso can be reformulated as a sequence of ℓ_1-penalized median regressions with two pseudo observations, to which existing packages for penalized QR can be applied. Moreover, Zheng, Peng and He [2018] used cross-validation to choose λ_0 (the initial penalty level) and the increment $c > 0$ by a two-dimensional grid search. In principle, we can apply this tuning scheme to both CQR-Lasso and its smoothed counterpart to achieve better variable selection performance. From a computational point of view, we apply a simpler tuning method by only choosing λ_0 via cross-validation and focus on speed comparisons. To be specific, we first compute cross-validated SCQR-Lasso and record its runtime, and then compute CQR-Lasso estimator using the same selected λ-sequence and record the runtime. For SCQR-Lasso, we apply the LAMM algorithm, described in Section C.1.2 of the Appendix, to compute each $\widehat{\boldsymbol{\beta}}(\tau_k)$ defined in (4.30); for CQR-Lasso, we use the `LASSO.fit` function in `rqPen` to fit the penalized median regression at each quantile level. The box plots of running time (in second) over 500 replications are displayed in Figure 4.5. On average, our implementation of the cross-validated SCQR-Lasso is more



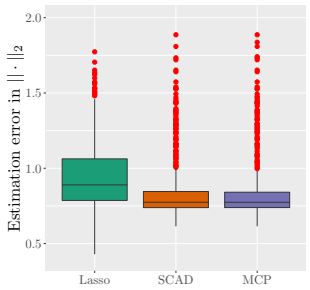
(a) FDR under homoscedastic model

(b) ℓ_2-error under homoscedastic model

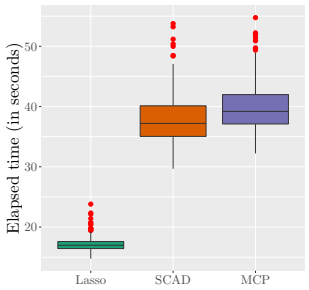
(c) Runtime under homoscedastic model



(d) FDR under heteroscedastic model



(e) ℓ_2-error under heteroscedastic model



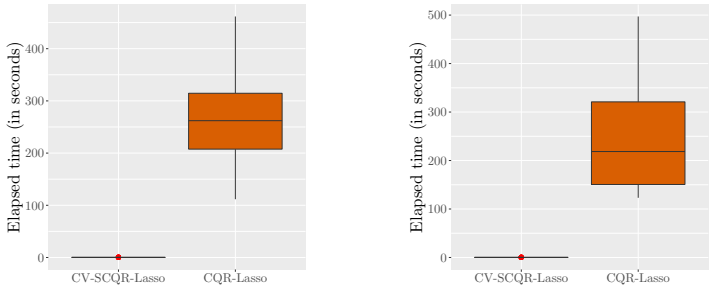
(f) Runtime under heteroscedastic model

Figure 4.4. Box plots of the false discovery rate, ℓ_2-error, and runtime for the ℓ_1, SCAD, and MCP regularized smoothed CQR. The true positive rates (TPR) are not visually informative, and thus are reported as follows. For the homoscedastic model, the average TPR are 1.00 for Lasso, 0.9996 for SCAD, and 0.9992 for MCP; for the heteroscedastic model, the average TPR are 0.9872 for Lasso, 0.919 for SCAD, and 0.917 for MCP. The censoring rates vary between 25% and 30%.

than 10 times faster than the CQR-Lasso implementation without cross-validation (18 seconds versus 250 seconds). We refer to He et al. [2022] for more tails of simulations. The code for the proposed method and our implementation of Zheng, Peng and He [2018]'s method is available at https://github.com/XiaoouPan/scqr.

4.6 Acknowledgements

This chapter, in part, is a reprint of the material in the paper "Scalable estimation and inference for censored quantile regression process", He, Xuming; Pan, Xiaoou; Tan, Kean Ming and Zhou, Wen-Xin. The paper has been submitted and is currently under major revision, 2022.



(a) Runtime under homoscedastic model

(b) Runtime under heteroscedastic model

Figure 4.5. Box plots of runtime for ℓ_1-penalized CQR (CQR-Lasso) and cross-validated ℓ_1-penalized smoothed CQR (CV-SCQR-Lasso). The censoring rates vary from 25% to 30%. CQR-Lasso is implemented using the `LASSO.fit` function in the package `rqPen`.

The dissertation author was the primary investigator and author of this paper.

Appendix A

Supplementary Material for Chapter 2

A.1 Proofs of Main Results

All the probabilistic bounds presented in the proof are non-asymptotic with explicit errors. The values of the constants involved are obtained with the goal of making the proof transparent, and may be improved by more careful calculations or under less general distributional assumptions on the covariates and noise variables.

A.1.1 Preliminaries

Recall that $Q_n(\boldsymbol{\beta}) = (1/n)\sum_{i=1}^n \rho_\tau(y_i - \langle \boldsymbol{x}_i, \boldsymbol{\beta}\rangle)$ is the empirical quantile loss function. Since $Q_n : \mathbb{R}^p \to \mathbb{R}$ is convex, we define its subdifferential ∂Q_n by

$$\partial Q_n(\boldsymbol{\beta}) = \big\{\boldsymbol{\xi} \in \mathbb{R}^p : Q_n(\boldsymbol{\beta}') \geq Q_n(\boldsymbol{\beta}) + \langle \boldsymbol{\xi}, \boldsymbol{\beta}' - \boldsymbol{\beta}\rangle \ \text{ for all } \ \boldsymbol{\beta} \in \mathbb{R}^p\big\}. \tag{A.1}$$

A vector $\boldsymbol{\xi} \in \partial Q_n(\boldsymbol{\beta})$ is called a subgradient of Q_n in $\boldsymbol{\beta}$. More specifically, the subdifferential ∂Q_n is the collection of vectors $\boldsymbol{\xi}_{\boldsymbol{\beta}} = (\xi_{\boldsymbol{\beta},1}, \ldots, \xi_{\boldsymbol{\beta},p})^\intercal$ satisfying, for $j = 1 \ldots, p$,

$$\begin{aligned}\xi_{\boldsymbol{\beta},j} = &-\frac{\tau}{n}\sum_{i=1}^n x_{ij} I(y_i > \langle \boldsymbol{x}_i, \boldsymbol{\beta}\rangle) \\ &+ \frac{1-\tau}{n}\sum_{i=1}^n x_{ij} I(y_i < \langle \boldsymbol{x}_i, \boldsymbol{\beta}\rangle) - \frac{1}{n}\sum_{i=1}^n x_{ij} v_i I(y_i = \langle \boldsymbol{x}_i, \boldsymbol{\beta}\rangle),\end{aligned} \tag{A.2}$$

where $v_i \in [\tau - 1, \tau]$.

Of particular interest is the subdifferential $\partial Q_n(\boldsymbol{\beta}^*)$ under model (2.1). By (A.2), every vector $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_p)^{\mathrm{T}} \in \partial Q_n(\boldsymbol{\beta}^*)$ can be written as

$$\begin{aligned}\xi_j = &-\frac{\tau}{n}\sum_{i=1}^{n} x_{ij}\{I(\varepsilon_i > 0) - (1-\tau)\} \\ &+ \frac{1-\tau}{n}\sum_{i=1}^{n} x_{ij}\{I(\varepsilon_i < 0) - \tau\} - \frac{1}{n}\sum_{i=1}^{n} x_{ij} v_i I(\varepsilon_i = 0), \quad j = 1, \ldots, p, \end{aligned} \tag{A.3}$$

where $v_i \in [\tau - 1, \tau]$.

Proposition A.1.1. *Assume Conditions 2.1.1 and 2.1.2 hold. Then, every subgradient $\boldsymbol{\xi}_{\boldsymbol{\beta}^*} \in \partial Q_n(\boldsymbol{\beta}^*)$ satisfies*

$$\mathbb{P}\left(\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\xi}_{\boldsymbol{\beta}^*}\|_2 \geq 3\upsilon_0\sqrt{\frac{2p+x}{n}} \right) \leq e^{-x}, \quad \textit{valid for any } x \geq 0.$$

The following proposition provides a form of the restricted convexity for the empirical quantile loss function.

Proposition A.1.2. *Assume Conditions 2.1.1 and 2.1.2 hold. Then, for any $t \geq 0$, it holds with probability at least $1 - e^{-t}/2$ that*

$$\langle \boldsymbol{\xi}_{\boldsymbol{\beta}} - \boldsymbol{\xi}_{\boldsymbol{\beta}^*}, \boldsymbol{\beta} - \boldsymbol{\beta}^* \rangle \geq \frac{1}{8}\underline{f}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma}}^2 - 4\upsilon_0^2\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma}}\sqrt{\frac{2(p+t)}{n}} \tag{A.4}$$

uniformly over $\boldsymbol{\beta} \in \mathbb{R}^p$ satisfying $0 \leq \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma}} \leq \underline{f}/(6L_0\upsilon_0^2)$.

Propositions A.1.1 and A.1.2 provide the key ingredients to prove Theorems 2.1.1 and 2.1.2. Similarly, the finite sample performance of the multiplier bootstrap estimator relies on the corresponding properties of the weighted quantile loss function, which are given by Propositions A.1.3 and A.1.4 below.

Recall that $\mathbb{P}^*$ and $\mathbb{E}^*$ denote, respectively, the probability measure and expectation (over

$\mathscr{R}_n = \{e_i\}_{i=1}^n$) conditioning on $\mathscr{D}_n = \{(y_i, \boldsymbol{x}_i)\}_{i=1}^n$. For $i = 1, \ldots, n$, define

$$\zeta_i = I(\varepsilon_i \leq 0) - \tau \quad \text{and} \quad \boldsymbol{z}_i = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}_i, \tag{A.5}$$

which satisfy $\mathbb{E}(\zeta_i|\boldsymbol{x}_i) = 0$, $\mathbb{E}(\zeta_i^2|\boldsymbol{x}_i) = \tau(1-\tau)$ and $\mathbb{E}(\boldsymbol{z}_i\boldsymbol{z}_i^{\mathrm{T}}) = \mathbf{I}_p$.

Proposition A.1.3. *Assume Conditions 2.1.1 and 2.1.2 hold, and let $\boldsymbol{\xi}^\flat \in \partial Q_n^\flat(\boldsymbol{\beta}^*)$. For any $t > 0$, there exists some event $\mathscr{G}_1(t) = \mathscr{G}_1(t; \mathscr{D}_n)$ with $\mathbb{P}\{\mathscr{G}_1(t)\} \geq 1 - e^{-2t}$ such that, with $\mathbb{P}^*$-probability at least $1 - e^{-2t}$ conditioned on $\mathscr{G}_1(t)$,*

$$\|\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\xi}^\flat - \mathbb{E}^*\boldsymbol{\xi}^\flat)\|_2 \leq 2\sqrt{\frac{p+t}{n}} \tag{A.6}$$

as long as $n \gtrsim p + t$.

Similarly to Proposition A.1.2, the following result establishes the restricted strong convexity for the weighted quantile loss function.

Proposition A.1.4. *Assume Conditions 2.1.1 and 2.1.2 hold. For any $t \geq 0$, there exists some event $\mathscr{G}_2(t) = \mathscr{G}_2(t; \mathscr{D}_n)$ such that $\mathbb{P}\{\mathscr{G}_2(t)\} \geq 1 - e^{-t}$, and with $\mathbb{P}^*$-probability at least $1 - e^{-t}/2$ conditioned on $\mathscr{G}_2(t)$,*

$$\langle \boldsymbol{\xi}_{\boldsymbol{\beta}}^\flat - \boldsymbol{\xi}_{\boldsymbol{\beta}^*}^\flat, \boldsymbol{\beta} - \boldsymbol{\beta}^* \rangle \geq \frac{1}{8}\underline{f}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma}}^2 - 8\upsilon_0^2\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma}}\sqrt{\frac{2(p+t)}{n}} \tag{A.7}$$

uniformly over $\boldsymbol{\beta} \in \mathbb{R}^p$ satisfying $0 \leq \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma}} \leq \underline{f}/(6L_0\upsilon_0^2)$ as long as $n \gtrsim \log(p) + t$.

The theory of classical quantile regression is not the focus of this dissertation, and we refer to Pan and Zhou [2021] for the proofs of Propositions A.1.1–A.1.4.

A.1.2 Proof of Theorem 2.1.1

By the convexity of $\boldsymbol{\beta} \mapsto Q_n(\boldsymbol{\beta})$, $\widehat{\boldsymbol{\beta}}$ satisfies the first-order condition that $\boldsymbol{\xi}_{\widehat{\boldsymbol{\beta}}} = \mathbf{0}$ for some $\boldsymbol{\xi}_{\widehat{\boldsymbol{\beta}}} \in \partial Q_n(\widehat{\boldsymbol{\beta}})$. The proof builds on the symmetrized Bregman divergence associated with

Q_n, defined as

$$D(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = \langle \boldsymbol{\xi}_{\boldsymbol{\beta}_1} - \boldsymbol{\xi}_{\boldsymbol{\beta}_2}, \boldsymbol{\beta}_1 - \boldsymbol{\beta}_2 \rangle, \text{ for } \boldsymbol{\xi}_{\boldsymbol{\beta}_1} \in \partial Q_n(\boldsymbol{\beta}_1), \boldsymbol{\xi}_{\boldsymbol{\beta}_2} \in \partial Q_n(\boldsymbol{\beta}_2).$$

By convexity, $D(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) \geq 0$ for any subdifferentials $\boldsymbol{\xi}_{\boldsymbol{\beta}_1}$ and $\boldsymbol{\xi}_{\boldsymbol{\beta}_2}$. Taking $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = (\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}^*)$, we have

$$0 \leq \langle \boldsymbol{\xi}_{\widehat{\boldsymbol{\beta}}} - \boldsymbol{\xi}_{\boldsymbol{\beta}^*}, \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \rangle = \langle -\boldsymbol{\xi}_{\boldsymbol{\beta}^*}, \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \rangle \leq \|\boldsymbol{\Sigma}^{-1/2} \boldsymbol{\xi}_{\boldsymbol{\beta}^*}\|_2 \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma}}, \tag{A.8}$$

for any $\boldsymbol{\xi}_{\boldsymbol{\beta}^*} \in \partial Q_n(\boldsymbol{\beta}^*)$. Starting with (A.8), we bound the left- and right-hand sides of (A.9) separately. To establish the lower bound, we use a localized argument [Sun, Zhou and Fan, 2020] and a new restricted strong convexity property for the empirical quantile loss (Proposition A.1.2).

Define the rescaled ℓ_2-ball $\mathbb{B}_{\boldsymbol{\Sigma}}(t) = \{\boldsymbol{\beta} \in \mathbb{R}^p : \|\boldsymbol{\beta}\|_{\boldsymbol{\Sigma}} \leq t\}$, $t > 0$. For some $0 < r \leq \underline{f}/(6L_0 \upsilon_0^2)$ to be determined, define

$$\eta = \sup\big\{u \in [0,1] : u(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \in \mathbb{B}_{\boldsymbol{\Sigma}}(r)\big\} \quad \text{and} \quad \widetilde{\boldsymbol{\beta}} = \boldsymbol{\beta}^* + \eta(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*).$$

By the above definition, $\eta = 1$ if $\widehat{\boldsymbol{\beta}} \in \boldsymbol{\beta}^* + \mathbb{B}_{\boldsymbol{\Sigma}}(r)$, and $\eta < 1$ if $\widehat{\boldsymbol{\beta}} \notin \boldsymbol{\beta}^* + \mathbb{B}_{\boldsymbol{\Sigma}}(r)$. In the latter case, we have $\widetilde{\boldsymbol{\beta}} \in \boldsymbol{\beta}^* + \partial \mathbb{B}_{\boldsymbol{\Sigma}}(r)$. Applying Lemma C.1 in Sun, Zhou and Fan [2020] with slight modifications yields the bound $D(\widetilde{\boldsymbol{\beta}}, \boldsymbol{\beta}^*) \leq \eta D(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}^*)$, leading to

$$\langle \boldsymbol{\xi}_{\widetilde{\boldsymbol{\beta}}} - \boldsymbol{\xi}_{\boldsymbol{\beta}^*}, \widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \rangle \leq \eta \langle \boldsymbol{\xi}_{\widehat{\boldsymbol{\beta}}} - \boldsymbol{\xi}_{\boldsymbol{\beta}^*}, \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \rangle, \tag{A.9}$$

where $\boldsymbol{\xi}_{\boldsymbol{\beta}^*} \in \partial Q_n(\boldsymbol{\beta}^*)$ and $\boldsymbol{\xi}_{\widetilde{\boldsymbol{\beta}}} \in \partial Q_n(\widetilde{\boldsymbol{\beta}})$. This, together with the fact $\boldsymbol{\xi}_{\widehat{\boldsymbol{\beta}}} = \mathbf{0}$ and Cauchy-Schwarz inequality, implies

$$\langle \boldsymbol{\xi}_{\widetilde{\boldsymbol{\beta}}} - \boldsymbol{\xi}_{\boldsymbol{\beta}^*}, \widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \rangle \leq \eta \langle -\boldsymbol{\xi}_{\boldsymbol{\beta}^*}, \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \rangle \leq \|\boldsymbol{\Sigma}^{-1/2} \boldsymbol{\xi}_{\boldsymbol{\beta}^*}\|_2 \|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma}}. \tag{A.10}$$

Note that (A.10) is a localized version of (A.8) because $\widetilde{\boldsymbol{\beta}}$ falls in a local neighborhood of $\boldsymbol{\beta}^*$.

Setting $\widetilde{\boldsymbol{\delta}} = \widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \in \mathbb{B}_{\boldsymbol{\Sigma}}(r)$, it follows from Proposition A.1.2 that

$$\langle \boldsymbol{\xi}_{\widetilde{\boldsymbol{\beta}}} - \boldsymbol{\xi}_{\boldsymbol{\beta}^*}, \widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \rangle \geq \frac{1}{8} \underline{f} \|\widetilde{\boldsymbol{\delta}}\|_{\boldsymbol{\Sigma}}^2 - 4\upsilon_0^2 \|\widetilde{\boldsymbol{\delta}}\|_{\boldsymbol{\Sigma}} \sqrt{\frac{2(p+t)}{n}}$$

with probability at least $1 - e^{-t}/2$. Combining this with (A.9) and (A.10), and taking $x = t > 0$ in Proposition A.1.1, we obtain

$$\frac{1}{8} \underline{f} \|\widetilde{\boldsymbol{\delta}}\|_{\boldsymbol{\Sigma}}^2 < \big(4\upsilon_0^2 + 3\upsilon_0\big) \|\widetilde{\boldsymbol{\delta}}\|_{\boldsymbol{\Sigma}} \sqrt{\frac{2(p+t)}{n}}$$

with probability at least $1 - 2e^{-t}$. Canceling $\|\widetilde{\boldsymbol{\delta}}\|_{\boldsymbol{\Sigma}}$ on both sides yields

$$\|\widetilde{\boldsymbol{\delta}}\|_{\boldsymbol{\Sigma}} < r := 8\underline{f}^{-1}(4\upsilon_0^2 + 3\upsilon_0) \sqrt{\frac{2(p+t)}{n}}$$

with probability at least $1 - 2e^{-t}$ as long as $n \geq CL_0^2 \underline{f}^{-4}(p+t)$ for some constant $C > 0$ depending only on υ_0. Consequently, $\widetilde{\boldsymbol{\beta}}$ falls in the interior of $\boldsymbol{\beta}^* + \mathbb{B}_{\boldsymbol{\Sigma}}(r)$, enforcing $\eta = 1$ and $\widehat{\boldsymbol{\beta}} = \widetilde{\boldsymbol{\beta}} \in \boldsymbol{\beta}^* + \mathbb{B}_{\boldsymbol{\Sigma}}(r)$. Otherwise if $\widehat{\boldsymbol{\beta}} \notin \boldsymbol{\beta}^* + \mathbb{B}_{\boldsymbol{\Sigma}}(r)$, we must have $\widetilde{\boldsymbol{\beta}}$ on the boundary, i.e. $\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma}} = r$, which leads to contradiction. This completes the proof. □

A.1.3 Proof of Theorem 2.1.2

To begin with, define the "gradient" function $\nabla Q_n : \mathbb{R}^p \to \mathbb{R}^p$ as

$$\nabla Q_n(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^n \boldsymbol{x}_i \big\{ I(y_i \leq \langle \boldsymbol{x}_i, \boldsymbol{\beta} \rangle) - \tau \big\}, \ \ \boldsymbol{\beta} \in \mathbb{R}^p. \tag{A.11}$$

Recall from Condition 2.1.2, the conditional distribution of ε given $\boldsymbol{x}$ is continuous. Lemma A.1 in Ruppert and Carroll [1980] states that with probability one, $\boldsymbol{\xi}_{\boldsymbol{\beta}} = \nabla Q_n(\boldsymbol{\beta})$ for any $\boldsymbol{\xi}_{\boldsymbol{\beta}} \in \partial Q_n(\boldsymbol{\beta})$. Hence, we will treat ∇Q_n as the gradient of Q_n throughout the proof. Moreover, consider the population loss $\mathbb{E}Q_n(\boldsymbol{\beta}) = \mathbb{E}\rho_\tau(y - \langle \boldsymbol{x}, \boldsymbol{\beta} \rangle)$, whose gradient vector and Hessian

matrix are given, respectively, by

$$\nabla \mathbb{E} Q_n(\boldsymbol{\beta}) = \mathbb{E}\big[\boldsymbol{x}\{I(\varepsilon \le \langle \boldsymbol{x}, \boldsymbol{\beta} - \boldsymbol{\beta}^* \rangle) - \tau\}\big] \text{ and } \nabla^2 \mathbb{E} Q_n(\boldsymbol{\beta}) = \mathbb{E}\big\{ f_{\varepsilon|\boldsymbol{x}}(\langle \boldsymbol{x}, \boldsymbol{\beta} - \boldsymbol{\beta}^* \rangle) \boldsymbol{x}\boldsymbol{x}^{\mathrm{T}} \big\}.$$

Next, define the vector-valued random process

$$\Delta(\boldsymbol{\beta}) = \mathbf{S}^{-1/2}\big\{\nabla Q_n(\boldsymbol{\beta}) - \nabla Q_n(\boldsymbol{\beta}^*)\big\} - \mathbf{S}^{1/2}(\boldsymbol{\beta} - \boldsymbol{\beta}^*), \tag{A.12}$$

where $\mathbf{S} = \nabla^2 \mathbb{E} Q_n(\boldsymbol{\beta}^*) = \mathbb{E}\{f_{\varepsilon|\boldsymbol{x}}(0)\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\}$. The goal is to bound $\|\Delta(\boldsymbol{\beta})\|_2$ uniformly over $\boldsymbol{\beta}$ in a local neighborhood of $\boldsymbol{\beta}^*$. To this end, we deal with $\mathbb{E}\Delta(\boldsymbol{\beta})$ and $\Delta(\boldsymbol{\beta}) - \mathbb{E}\Delta(\boldsymbol{\beta})$ separately, starting with $\mathbb{E}\Delta(\boldsymbol{\beta})$. Applying the mean value theorem for vector-valued functions yields

$$\begin{aligned} \mathbb{E}\Delta(\boldsymbol{\beta}) &= \mathbf{S}^{-1/2}\left\langle \int_0^1 \nabla^2 \mathbb{E} Q_n(\boldsymbol{\beta}_t^*)\mathrm{d}t, \boldsymbol{\beta} - \boldsymbol{\beta}^* \right\rangle - \mathbf{S}^{1/2}(\boldsymbol{\beta} - \boldsymbol{\beta}^*) \\ &= \left\langle \mathbf{S}^{-1/2} \int_0^1 \nabla^2 \mathbb{E} Q_n(\boldsymbol{\beta}_t^*)\mathrm{d}t\, \mathbf{S}^{-1/2} - \mathbf{I}_p, \mathbf{S}^{1/2}(\boldsymbol{\beta} - \boldsymbol{\beta}^*) \right\rangle, \end{aligned} \tag{A.13}$$

where $\boldsymbol{\beta}_t^* = (1-t)\boldsymbol{\beta}^* + t\boldsymbol{\beta}$ and $\nabla^2 \mathbb{E} Q_n(\boldsymbol{\beta}_t^*) = \mathbb{E}\{f_{\varepsilon|\boldsymbol{x}}(t\langle \boldsymbol{x}, \boldsymbol{\beta} - \boldsymbol{\beta}^* \rangle)\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\}$. For $r > 0$, define the local elliptic neighborhood of $\boldsymbol{\beta}^*$ as $\Theta_{\boldsymbol{\Sigma}}(r) := \big\{\boldsymbol{\beta} \in \mathbb{R}^p : \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma}} \le r\big\}$. By Conditions 2.1.1 and 2.1.2, $\boldsymbol{\Sigma}$ is positive definite and $\underline{f} \le f_{\varepsilon|\boldsymbol{x}}(0) \le \bar{f}$, so that $\underline{f}\boldsymbol{\Sigma} \preceq \mathbf{S} \preceq \bar{f}\boldsymbol{\Sigma}$. For $\boldsymbol{\delta} = \boldsymbol{\beta} - \boldsymbol{\beta}^*$ with $\boldsymbol{\beta} \in \Theta_{\boldsymbol{\Sigma}}(r)$, the Lipschitz continuity of $f_{\varepsilon|\boldsymbol{x}}$ ensures that

$$\begin{aligned} &\big\|\mathbf{S}^{-1/2}\nabla^2 \mathbb{E} Q_n(\boldsymbol{\beta}_t^*)\mathbf{S}^{-1/2} - \mathbf{I}_p\big\|_2 = \big\|\mathbf{S}^{-1/2}\mathbb{E}\big[\big\{f_{\varepsilon|\boldsymbol{x}}(t\langle \boldsymbol{x}, \boldsymbol{\delta} \rangle) - f_{\varepsilon|\boldsymbol{x}}(0)\big\}\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\big]\mathbf{S}^{-1/2}\big\|_2 \\ &\le L_0 t \cdot \sup_{\boldsymbol{u} \in \mathbb{B}^p(1)} \mathbb{E}\big\{\langle \mathbf{S}^{-1/2}\boldsymbol{x}, \boldsymbol{u} \rangle^2 |\langle \boldsymbol{x}, \boldsymbol{\delta} \rangle|\big\} \\ &\le \underline{f}^{-1} L_0 t \cdot \left(\sup_{\boldsymbol{u} \in \mathbb{B}^p(1)} \mathbb{E}|\langle \boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}, \boldsymbol{u} \rangle|^3 \right)^{2/3} \big(\mathbb{E}|\langle \boldsymbol{x}, \boldsymbol{\delta} \rangle|^3\big)^{1/3} \\ &\le L_0 \underline{f}^{-1} m_3 r t, \end{aligned}$$

where $m_k := \sup_{\boldsymbol{u} \in \mathbb{B}^p(1)} \mathbb{E}|\langle \boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}, \boldsymbol{u} \rangle|^k$ (for $k \ge 1$) depends only on υ_0 and k. Combining this

with (A.13), we obtain

$$\sup_{\boldsymbol{\beta}\in\Theta_{\boldsymbol{\Sigma}}(r)} \|\mathbb{E}\Delta(\boldsymbol{\beta})\|_2 \le \frac{1}{2} L_0 \underline{f}^{-1} \bar{f}^{1/2} m_3 r^2. \tag{A.14}$$

Turning to the stochastic term $\Delta(\boldsymbol{\beta}) - \mathbb{E}\Delta(\boldsymbol{\beta})$, define the centered gradient function

$$R_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} (1-\mathbb{E})\big\{I(\langle \boldsymbol{x}_i, \boldsymbol{\beta}-\boldsymbol{\beta}^*\rangle \ge \varepsilon_i) - \tau\big\}\boldsymbol{x}_i,$$

so that $\Delta(\boldsymbol{\beta}) - \mathbb{E}\Delta(\boldsymbol{\beta}) = \mathbf{S}^{-1/2}\{R_n(\boldsymbol{\beta}) - R_n(\boldsymbol{\beta}^*)\}$. By a change of variable $\boldsymbol{v} = \boldsymbol{\Sigma}^{1/2}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)$, we have

$$\begin{aligned}
\sup_{\boldsymbol{\beta}\in\Theta_{\boldsymbol{\Sigma}}(r)} \|\Delta(\boldsymbol{\beta}) - \mathbb{E}\Delta(\boldsymbol{\beta})\|_2 &\le \underline{f}^{-1/2} \sup_{\boldsymbol{\beta}\in\Theta_{\boldsymbol{\Sigma}}(r)} \|\boldsymbol{\Sigma}^{-1/2}\{R_n(\boldsymbol{\beta}) - R_n(\boldsymbol{\beta}^*)\}\|_2 \\
&= \underline{f}^{-1/2} \sup_{\boldsymbol{v}\in\mathbb{B}^p(r)} \|\boldsymbol{\Sigma}^{-1/2}\{R_n(\boldsymbol{\beta}^* + \boldsymbol{\Sigma}^{-1/2}\boldsymbol{v}) - R_n(\boldsymbol{\beta}^*)\}\|_2 \\
&= \underline{f}^{-1/2} r^{-1} \sup_{\boldsymbol{u},\boldsymbol{v}\in\mathbb{B}^p(r)} \underbrace{\langle \boldsymbol{\Sigma}^{-1/2}\{R_n(\boldsymbol{\beta}^* + \boldsymbol{\Sigma}^{-1/2}\boldsymbol{v}) - R_n(\boldsymbol{\beta}^*)\}, \boldsymbol{u}\rangle}_{n^{-1/2}\Delta_0(\boldsymbol{u},\boldsymbol{v})},
\end{aligned} \tag{A.15}$$

where $\Delta_0(\boldsymbol{u},\boldsymbol{v}) = n^{-1/2}\sum_{i=1}^n (1-\mathbb{E})\langle \boldsymbol{z}_i, \boldsymbol{u}\rangle\{I(\varepsilon_i \le \langle \boldsymbol{z}_i, \boldsymbol{v}\rangle) - I(\varepsilon_i \le 0)\}$. To bound its supremum, we first show its concentration around the mean, and then bound the mean via a maximal inequality specialized to VC type classes (see, e.g., Chapter 2.6 in van der Vaart and Wellner [1996]). Consider the following two classes of real-valued functions on $\mathbb{R}\times\mathbb{R}^p$:

$$\mathscr{F}_1 = \{(z_0, \boldsymbol{z}) \mapsto \langle \boldsymbol{z}, \boldsymbol{u}\rangle : \boldsymbol{u} \in \mathbb{B}^p(r)\} \;\text{ and }\; \mathscr{F}_2 = \{(z_0, \boldsymbol{z}) \mapsto I(\langle \boldsymbol{z}, \boldsymbol{v}\rangle - z_0 \ge 0) : \boldsymbol{v} \in \mathbb{B}^p(r)\}. \tag{A.16}$$

Moreover, define the function $f_0 : (z_0, \boldsymbol{z}) \mapsto I(z_0 \le 0)$, and write $\bar{z}_i = (\varepsilon_i, \boldsymbol{z}_i) \in \mathbb{R}\times\mathbb{R}^p$ for $i = 1, \ldots, n$. Then, the supremum $\sup_{\boldsymbol{u},\boldsymbol{v}\in\mathbb{B}^p(r)} \Delta_0(\boldsymbol{u},\boldsymbol{v})$ can be written as the supremum of an

empirical process:

$$\sup_{\boldsymbol{v},\boldsymbol{u}\in\mathbb{B}^p(r)} \Delta_0(\boldsymbol{u},\boldsymbol{v}) = \sup_{f\in\mathscr{F}} \underbrace{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{f(\bar{z}_i)-\mathbb{E}f(\bar{z}_i)\}}_{\mathbb{G}_n f}, \tag{A.17}$$

where $\mathscr{F} = \mathscr{F}_1 \cdot (\mathscr{F}_2 - f_0)$ is the pointwise product of $\mathscr{F}_1$ and $\mathscr{F}_2 - f_0$. Under the assumption that $\sup_u |f_{\varepsilon|\boldsymbol{x}}(u)| \le M_0$ almost surely, we have, for each $i \in [n]$, $\sup_{f\in\mathscr{F}} f(\bar{\boldsymbol{z}}_i) \le r\|\boldsymbol{z}_i\|_2$ and $\sup_{f\in\mathscr{F}} \mathbb{E}f(\bar{\boldsymbol{z}}_i)^2 \le M_0 \sup_{\boldsymbol{u},\boldsymbol{v}\in\mathbb{B}^p(r)} \mathbb{E}\langle \boldsymbol{z}_i,\boldsymbol{u}\rangle^2 |\langle \boldsymbol{z}_i,\boldsymbol{v}\rangle| \le M_0 m_3 r^3$. By Lemma 2.2.2 in van der Vaart and Wellner [1996],

$$\begin{aligned}\left\| \max_{1\le i\le n} \sup_{f\in\mathscr{F}} |f(\bar{\boldsymbol{z}}_i)| \right\|_{\psi_1} &\le r \left\| \max_{1\le i\le n} \|\boldsymbol{z}_i\|_2 \right\|_{\psi_1} \le r p^{1/2} \left\| \max_{1\le i\le n, 1\le j\le p} |z_{ij}| \right\|_{\psi_1} \\ &\le (\log 2)^{1/2} r p^{1/2} \left\| \max_{1\le i\le n, 1\le j\le p} |z_{ij}| \right\|_{\psi_2} \le c_0 (p\log n)^{1/2} r,\end{aligned}$$

where $c_0 > 0$ depends only on υ_0, and $\|\cdot\|_{\psi_q}$ $(1 \le q \le 2)$ denotes the ψ_q-Orlicz norm. Applying Theorem 4 in Adamczak [2008] with $\alpha = 1$ and $\delta = \eta = 1/2$, we obtain that for any $x \ge 0$,

$$\sup_{f\in\mathscr{F}} \mathbb{G}_n f \le \frac{3}{2}\mathbb{E}\left(\sup_{f\in\mathscr{F}} \mathbb{G}_n f \right) + x$$

with probability at least $1 - e^{-x^2/(3M_0 m_3 r^3)} - 3e^{-x\sqrt{n}/\{c_1 (p\log n)^{1/2} r\}}$, where $c_1 > 0$ depends only on c_0. Given $t \ge 0$ such that $4e^{-t} \le 1$, taking

$$x = \max\left\{ (3M_0 m_3)^{1/2} r^{3/2} t^{1/2}, 2c_1 r t (p\log n)^{1/2} n^{-1/2} \right\}$$

in the above bound yields that, with probability at least $1 - e^{-t} - 3e^{-2t} \ge 1 - 2e^{-t}$,

$$\sup_{f\in\mathscr{F}} \mathbb{G}_n f \le \frac{3}{2}\mathbb{E}\left(\sup_{f\in\mathscr{F}} \mathbb{G}_n f \right) + \max\left\{ (3M_0 m_3)^{1/2} r^{3/2} t^{1/2}, 2c_1 r t \sqrt{\frac{p\log n}{n}} \right\}. \tag{A.18}$$

To bound $\mathbb{E}(\sup_{f\in\mathscr{F}} \mathbb{G}_n f)$, we control the covering numbers $N(\mathscr{F}, L_2(Q), \varepsilon\|F\|_{Q,2})$ for

all finitely supported probability measures Q on $\mathbb{R}\times\mathbb{R}^p$ and $0<\varepsilon<1$, where $F(\bar{\boldsymbol{z}})=r\|\boldsymbol{z}\|_2$ is a measurable envelope of $\mathscr{F}$. Respectively, for the function classes $\mathscr{F}_1$ and $\mathscr{F}_2$ that have envelopes $F_1(\bar{\boldsymbol{z}})=r\|\boldsymbol{z}\|_2$ and $F_2(\bar{\boldsymbol{z}})=1$, using Theorem B in Dudley [1979] and Theorem 2.6.7 in van der Vaart and Wellner [1996] we have

$$\sup_Q N(\mathscr{F}_1,L_2(Q),\varepsilon\|F_1\|_{Q,2})\le(A_1/\varepsilon)^{2(p+2)} \text{ and } \sup_Q N(\mathscr{F}_2,L_2(Q),\varepsilon)\le(A_1/\varepsilon)^{2(p+2)}$$

for some $A_1>e$, where the suprema are taken over all finitely discrete probability measures Q on $\mathbb{R}\times\mathbb{R}^p$. Combining the above bounds with Corollary A.1 in the supplement of Chernozhukov, Chetverikov and Kato [2014] shows that

$$\begin{aligned}&\sup_Q N(\mathscr{F},L_2(Q),\varepsilon\|F\|_{Q,2})\\&\le\sup_Q N(\mathscr{F}_1,L_2(Q),2^{-1/2}\varepsilon\|F_1\|_{Q,2})\cdot\sup_Q N(\mathscr{F}_2,L_2(Q),2^{-1/2}\varepsilon)\le(A_2/\varepsilon)^{4(p+2)},\end{aligned}$$

where $A_2=2^{1/2}A_1$. For the envelop function $F:\mathbb{R}\times\mathbb{R}^p\to\mathbb{R}^+$, we have $\mathbb{E}F(\boldsymbol{z})^2=r^2p$. Consequently, it follows from Corollary 5.1 in Chernozhukov, Chetverikov and Kato [2014] that

$$\mathbb{E}\left(\sup_{f\in\mathscr{F}}\mathbb{G}_nf\right)\lesssim\sqrt{M_0m_3r^3p\log\big(A_2^2p/(M_0m_3r)\big)}+rM_n\frac{p}{n^{1/2}}\log\big(A_2^2p/(M_0m_3r)\big),\qquad(A.19)$$

where $M_n:=(\mathbb{E}\max_{1\le i\le n}\|\boldsymbol{z}_i\|_2^2)^{1/2}$. To bound M_n, we will reply on an exponential-type tail inequality for $X:=\max_{1\le i\le n}\|\boldsymbol{z}_i\|_2^2$. Assume there exist constants $A,a>0$ such that $\mathbb{P}(X\ge A+au)\le e^{-u}$ for every $u\in\mathbb{R}$. Then

$$\begin{aligned}\mathbb{E}(X)&=\int_0^\infty\mathbb{P}(X\ge t)\mathrm{d}t\le A+\int_A^\infty\mathbb{P}(X\ge t)\mathrm{d}t\\&=A+\int_0^\infty\mathbb{P}(X\ge A+t)\mathrm{d}t=A+a\int_0^\infty\mathbb{P}(X\ge A+au)\mathrm{d}u\le A+a.\end{aligned}$$

Given $\varepsilon\in(0,1)$, there exits a finite subset $\mathscr{N}_\varepsilon\subseteq\mathbb{S}^{p-1}$ with $|\mathscr{N}_\varepsilon|\le(1+2/\varepsilon)^p$ such that

$\max_{1\le i\le n}\|\boldsymbol{z}_i\|_2 \le (1-\varepsilon)^{-1}\max_{1\le i\le n}\max_{\boldsymbol{u}\in\mathscr{N}_\varepsilon}\langle \boldsymbol{u},\boldsymbol{w}_i\rangle$. For every $i\in[n]$ and $\boldsymbol{u}\in\mathscr{N}_\varepsilon$, Condition 2.1.1 indicates that $\mathbb{P}(|\langle \boldsymbol{u},\boldsymbol{w}_i\rangle|\ge \upsilon_0 u)\le 2e^{-u^2/2}$ for any $u\in\mathbb{R}$. Taking the union bound over $i\in[n]$ and $\boldsymbol{u}\in\mathscr{N}_\varepsilon$, and setting $u=\sqrt{2v+2\log(2n)+2p\log(1+2/\varepsilon)}$ ($v>0$), we obtain that with probability at least $1-2n(1+2/\varepsilon)^p e^{-u^2/2}=1-e^{-v}$, $\max_{1\le i\le n}\|\boldsymbol{z}_i\|_2\le (1-\varepsilon)^{-1}\upsilon_0\sqrt{2v+2\log(2n)+2p\log(1+2/\varepsilon)}$. Minimizing this upper bound with respect to $\varepsilon\in(0,1)$, we conclude that

$$\mathbb{P}\bigg[\max_{1\le i\le n}\|\boldsymbol{z}_i\|_2^2\ge 2\upsilon_0^2\big\{3.7d+\log(2n)+v\big\}\bigg]\le e^{-v},\ \text{valid for every } v>0.$$

Taking $A=2\upsilon_0^2\{3.7p+\log(2n)\}$ and $a=2\upsilon_0^2$ in the earlier analysis yields the bound $M_n^2=\mathbb{E}(\max_{1\le i\le n}\|\boldsymbol{z}_i\|_2^2)\le 2\upsilon_0^2\{3.7p+\log(2en)\}$. Plugging this into (A.19) gives

$$\mathbb{E}\bigg(\sup_{f\in\mathscr{F}}\mathbb{G}_n f\bigg)\lesssim\sqrt{M_0m_3r^3p\log\big(A_2^2p/(M_0m_3r)\big)}+r(p+\log n)^{1/2}\frac{p}{n^{1/2}}\log\big(A_2^2p/(M_0m_3r)\big).\tag{A.20}$$

Together, (A.15), (A.17), (A.18) and (A.20) imply that with probability at least $1-2e^{-t}$,

$$\begin{aligned}&\sup_{\boldsymbol{\beta}\in\Theta_{\boldsymbol{\Sigma}}(r)}\|\Delta(\boldsymbol{\beta})-\mathbb{E}\Delta(\boldsymbol{\beta})\|_2\\&\le C_1\bigg\{\sqrt{\frac{rt}{n}}+\sqrt{\log(C_2p/r)\frac{rp}{n}}+(p+\log n)^{1/2}\log(C_2p/r)\frac{p}{n}+(p\log n)^{1/2}\frac{t}{n}\bigg\}.\end{aligned}\tag{A.21}$$

Thus far, we have established a high probability bound on the ℓ_2-norm of $\Delta(\boldsymbol{\beta})=\mathbf{S}^{-1/2}\big\{\nabla Q_n(\boldsymbol{\beta})-\nabla Q_n(\boldsymbol{\beta}^*)\big\}-\mathbf{S}^{1/2}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)$ uniformly over $\boldsymbol{\beta}\in\Theta_{\boldsymbol{\Sigma}}(r)$, a local neighborhood of $\boldsymbol{\beta}^*$, for any prespecified $r>0$. By Theorem 2.1.1, we have $\widehat{\boldsymbol{\beta}}\in\Theta_{\boldsymbol{\Sigma}}(r_t)$ with probability at least $1-2e^{-t}$ as long as $n\ge CL_0^2\underline{f}^{-4}(p+t)$, where $r_t=C_3\sqrt{(p+t)/n}$. Setting $r=r_t$ in (A.14)

and (A.21), we find that with probability at least $1-2e^{-t}$,

$$\sup_{\boldsymbol{\beta}\in\Theta_{\boldsymbol{\Sigma}}(r_t)} \|\Delta(\boldsymbol{\beta})\|_2 \lesssim \frac{(p+t)^{1/4}(p\log n+t)^{1/2}}{n^{3/4}} + (p+\log n)^{1/2}\frac{p\log n}{n} + (p\log n)^{1/2}\frac{t}{n}.$$

Recalling that $\nabla Q_n(\widehat{\boldsymbol{\beta}}) = \mathbf{0}$, this completes the proof. □

A.1.4 Proof of Theorem 2.1.3

Let $\boldsymbol{\lambda} \in \mathbb{R}^p$ be an arbitrary vector defining a linear contrast. Define the normalized partial sum $S_n = n^{-1/2}\sum_{i=1}^n \gamma_i\zeta_i$ of independent zero-mean random variables, where $\zeta_i = I(\varepsilon_i \le 0) - \tau$ and $\gamma_i = -\langle \mathbf{S}^{-1}\boldsymbol{\lambda}, \boldsymbol{x}_i\rangle$. Moreover, write $\delta_n = (p+\log n)^{1/4}(p\log n)^{1/2}n^{-1/4} + (p+\log n)^{1/2}p\log(n)\,n^{-1/2}$. Applying Theorem 2.1.2 with $t = \log n$ yields that, under the scaling $n \gtrsim p+\log n$,

$$\begin{aligned}
&\left|n^{1/2}\langle\boldsymbol{\lambda}, \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\rangle - S_n\right| \\
&= n^{1/2}\left|\left\langle \mathbf{S}^{-1/2}\boldsymbol{\lambda}, \mathbf{S}^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) + \mathbf{S}^{-1/2}\frac{1}{n}\sum_{i=1}^n \big\{I(\varepsilon_i \le 0) - \tau\big\}\boldsymbol{x}_i \right\rangle\right| \le c_1\|\mathbf{S}^{-1/2}\boldsymbol{\lambda}\|_2\,\delta_n \quad \text{(A.22)}
\end{aligned}$$

with probability at least $1-4n^{-1}$ for some constant $c_1 > 0$.

For the partial sum S_n, note that $\mathrm{var}(S_n) = \sigma_\tau^2 = \tau(1-\tau)\|\mathbf{S}^{-1}\boldsymbol{\lambda}\|_{\boldsymbol{\Sigma}}^2$. Then it follows from the Berry-Esseen inequality (see, e.g., Tyurin [2011]) that

$$\begin{aligned}
&\sup_{x\in\mathbb{R}}\left|\mathbb{P}\big\{S_n \le \mathrm{var}(S_n)^{1/2}x\big\} - \Phi(x)\right| \\
&\le \frac{\mathbb{E}|\{I(\varepsilon \le 0) - \tau\}\langle \mathbf{S}^{-1}\boldsymbol{\lambda}, \boldsymbol{x}\rangle|^3}{2n^{1/2}\sigma_\tau^3} \le \frac{1-2(\tau-\tau^2)}{2(\tau-\tau^2)^{1/2}}\frac{m_3}{n^{1/2}} = c_2 n^{-1/2}. \quad \text{(A.23)}
\end{aligned}$$

Moreover, for any $a \le b$, $\Phi(b/\sigma_\tau) - \Phi(a/\sigma_\tau) \le (2\pi)^{-1/2}(b-a)/\sigma_\tau$. Combining this with

(A.22) and (A.23), for any $x \in \mathbb{R}$, we obtain

$$\begin{aligned}
&\mathbb{P}\big(n^{1/2}\langle \boldsymbol{\lambda}, \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\rangle \le x\big) \\
&\le \mathbb{P}\big(S_n \le x + c_1 \|\mathbf{S}^{-1/2}\boldsymbol{\lambda}\|_2\, \delta_n\big) + 4n^{-1} \\
&\le \mathbb{P}\big\{\mathrm{var}(S_n)^{1/2} G \le x + c_1 \|\mathbf{S}^{-1/2}\boldsymbol{\lambda}\|_2\, \delta_n\big\} + c_2 n^{-1/2} + 4n^{-1} \\
&\le \mathbb{P}\big(\sigma_\tau G \le x\big) + c_1\{2\pi\tau(1-\tau)\}^{-1/2}\delta_n + c_2 n^{-1/2} + 4n^{-1},
\end{aligned}$$

where $G \sim \mathscr{N}(0,1)$. A similar argument leads to the reverse inequality. Putting together the pieces established the Berry-Esseen bound (2.6). □

A.1.5 Proof of Theorem 2.1.4

Without loss of generality, we assume $t > 0$ is such that $2e^{-t} \le 1$ throughout the proof. By the convexity of $\boldsymbol{\beta} \mapsto Q_n^\flat(\boldsymbol{\beta})$, $\widehat{\boldsymbol{\beta}}^\flat$ satisfies the first-order condition that $\boldsymbol{\xi}^\flat_{\widehat{\boldsymbol{\beta}}^\flat} = \mathbf{0}$ for some $\boldsymbol{\xi}^\flat_{\widehat{\boldsymbol{\beta}}^\flat} \in \partial Q_n^\flat(\widehat{\boldsymbol{\beta}}^\flat)$. Again, we follow the same localized analysis as in the proof of Theorem 2.1.1. For some $0 < r \le \underline{f}/(6L_0\upsilon_0^2)$ to be determined, if $\widehat{\boldsymbol{\beta}}^\flat \notin \boldsymbol{\beta}^* + \mathbb{B}_{\boldsymbol{\Sigma}}(r)$, there exists $\eta \in (0,1)$ such that $\widetilde{\boldsymbol{\beta}} := \boldsymbol{\beta}^* + \eta(\widehat{\boldsymbol{\beta}}^\flat - \boldsymbol{\beta}^*) \in \boldsymbol{\beta}^* + \partial\mathbb{B}_{\boldsymbol{\Sigma}}(r)$; otherwise if $\widehat{\boldsymbol{\beta}}^\flat \in \boldsymbol{\beta}^* + \mathbb{B}_{\boldsymbol{\Sigma}}(r)$, we take $\eta = 1$ so that $\widetilde{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}^\flat$.

Similar to (A.9) and (A.10), we have that for any $\boldsymbol{\xi}^\flat_{\boldsymbol{\beta}^*} \in \partial Q_n^\flat(\boldsymbol{\beta}^*)$ and $\boldsymbol{\xi}^\flat_{\widetilde{\boldsymbol{\beta}}} \in \partial Q_n^\flat(\widetilde{\boldsymbol{\beta}})$,

$$\langle \boldsymbol{\xi}^\flat_{\widetilde{\boldsymbol{\beta}}} - \boldsymbol{\xi}^\flat_{\boldsymbol{\beta}^*}, \widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\rangle \le \|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\xi}^\flat_{\boldsymbol{\beta}^*}\|_2 \|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma}}.$$

For the right-hand side, Proposition A.1.3 implies that there exists some event $\mathscr{G}_1(t)$ with $\mathbb{P}\{\mathscr{G}_1(t)\} \ge 1 - e^{-2t}$ such that, conditioned on $\mathscr{G}_1(t)$,

$$\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\xi}^\flat_{\boldsymbol{\beta}^*}\|_2 \le 2\sqrt{\frac{p+t}{n}} + \|\boldsymbol{\Sigma}^{-1/2}\mathbb{E}^*\boldsymbol{\xi}^\flat_{\boldsymbol{\beta}^*}\|_2$$

with $\mathbb{P}^*$-probability at least $1 - e^{-2t}$ as long as $n \gtrsim p + t$. On the other hand, since $\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma}} \le r$,

by Proposition A.1.4, there exists some event $\mathscr{G}_2(t) = \mathscr{G}_2(t; \mathscr{D}_n)$ with $\mathbb{P}\{\mathscr{G}_2(t)\} \geq 1 - e^{-t}$ such that, conditioned on $\mathscr{G}_2(t)$,

$$\langle \boldsymbol{\xi}^\flat_{\widetilde{\boldsymbol{\beta}}} - \boldsymbol{\xi}^\flat_{\boldsymbol{\beta}^*}, \widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \rangle \geq \frac{1}{8} \underline{f} \|\widetilde{\boldsymbol{\delta}}\|^2_{\boldsymbol{\Sigma}} - 8\upsilon_0^2 \|\widetilde{\boldsymbol{\delta}}\|_{\boldsymbol{\Sigma}} \sqrt{\frac{2(p+t)}{n}}$$

with $\mathbb{P}^*$-probability at least $1 - e^{-t}/2$ as long as $n \gtrsim \log(p) + t$, where $\widetilde{\boldsymbol{\delta}} = \widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$. Together, the last three displays imply

$$\|\widetilde{\boldsymbol{\delta}}\|_{\boldsymbol{\Sigma}} \leq 8\underline{f}^{-1}\big(2^{1/2} + 8\upsilon_0^2\big) \sqrt{\frac{2(p+t)}{n}} + 8\underline{f}^{-1} \|\boldsymbol{\Sigma}^{-1/2} \mathbb{E}^* \boldsymbol{\xi}^\flat_{\boldsymbol{\beta}^*}\|_2 \tag{A.24}$$

with $\mathbb{P}^*$-probability at least $1 - e^{-t}$ conditioned on $\mathscr{G}_1(t) \cap \mathscr{G}_2(t)$.

For $\|\boldsymbol{\Sigma}^{-1/2} \mathbb{E}^* \boldsymbol{\xi}^\flat_{\boldsymbol{\beta}^*}\|_2$, it follows from (A.3) and Proposition A.1.1 that

$$\|\boldsymbol{\Sigma}^{-1/2} \mathbb{E}^* \boldsymbol{\xi}^\flat_{\boldsymbol{\beta}^*}\|_2 < 3\upsilon_0 \sqrt{\frac{2(p+t)}{n}} \tag{A.25}$$

with probability at least $1 - e^{-2t}$. Let $\mathscr{G}_3(t)$ be the event that (A.25) holds so that $\mathbb{P}\{\mathscr{G}_3(t)\} \geq 1 - e^{-2t}$.

Combining (A.24) and (A.25), we conclude that conditioned on $\mathscr{G}_1(t) \cap \mathscr{G}_2(t) \cap \mathscr{G}_3(t)$, $\|\widetilde{\boldsymbol{\delta}}\|_{\boldsymbol{\Sigma}} < r := C_4 \underline{f}^{-1} \sqrt{(p+t)/n}$ with $\mathbb{P}^*$-probability at least $1 - e^{-t}$ as long as $n \geq C_5 L_0^2 \underline{f}^{-4}(p + t)$, and $\mathbb{P}\{\mathscr{G}_1(t) \cap \mathscr{G}_2(t) \cap \mathscr{G}_3(t)\} \geq 1 - 2e^{-t}$, where the constants $C_4, C_5 > 0$ depend only on υ_0. This enforces $\widetilde{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}^\flat$. Finally, taking $\mathscr{E}(t) = \mathscr{G}_1(t) \cap \mathscr{G}_2(t) \cap \mathscr{G}_3(t)$ establishes the claim. □

A.1.6 Proof of Theorem 2.1.5

Following the proof of Theorem 2.1.2, we treat $\nabla Q^\flat_n(\boldsymbol{\beta}) := (1/n) \sum_{i=1}^n w_i \boldsymbol{x}_i \{I(y_i \leq \langle \boldsymbol{x}_i, \boldsymbol{\beta} \rangle) - \tau\}$ as the gradient of $Q^\flat_n(\boldsymbol{\beta})$. Under this notation, define the vector-valued random process

$$\Delta^\flat(\boldsymbol{\beta}) = \mathbf{S}^{-1/2} \big\{ \nabla Q^\flat_n(\boldsymbol{\beta}) - \nabla Q^\flat_n(\boldsymbol{\beta}^*) \big\} - \mathbf{S}^{1/2}(\boldsymbol{\beta} - \boldsymbol{\beta}^*) \text{ for } \boldsymbol{\beta} \in \mathbb{R}^p.$$

Recalling $\mathbb{E}(w_i) = 1$, we have $\mathbb{E}^* \nabla Q_n^\flat(\boldsymbol{\beta}) = \nabla Q_n(\boldsymbol{\beta}) = (1/n)\sum_{i=1}^n \boldsymbol{x}_i\{I(y_i \le \langle \boldsymbol{x}_i, \boldsymbol{\beta}\rangle) - \tau\}$. Define $R_n^\flat(\boldsymbol{\beta}) = \nabla Q_n^\flat(\boldsymbol{\beta}) - \nabla Q_n(\boldsymbol{\beta})$, so that

$$\Delta^\flat(\boldsymbol{\beta}) = \mathbf{S}^{-1/2}\big\{R_n^\flat(\boldsymbol{\beta}) - R_n^\flat(\boldsymbol{\beta}^*) + \nabla Q_n(\boldsymbol{\beta}) - \nabla Q_n(\boldsymbol{\beta}^*) - \mathbf{S}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)\big\}$$

and $\mathbb{E}^*\Delta^\flat(\boldsymbol{\beta}) = \Delta(\boldsymbol{\beta})$ with $\Delta(\boldsymbol{\beta})$ defined in (A.12). By the triangle inequality, for any $r > 0$ we have

$$\sup_{\boldsymbol{\beta}\in\Theta_{\boldsymbol{\Sigma}}(r)} \|\Delta^\flat(\boldsymbol{\beta})\|_2 \le \sup_{\boldsymbol{\beta}\in\Theta_{\boldsymbol{\Sigma}}(r)} \|\Delta^\flat(\boldsymbol{\beta}) - \mathbb{E}^*\Delta^\flat(\boldsymbol{\beta})\|_2 + \sup_{\boldsymbol{\beta}\in\Theta_{\boldsymbol{\Sigma}}(r)} \|\Delta(\boldsymbol{\beta})\|_2, \tag{A.26}$$

where $\Theta_{\boldsymbol{\Sigma}}(r) = \{\boldsymbol{\beta} \in \mathbb{R}^p : \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma}} \le r\}$.

The last term $\sup_{\boldsymbol{\beta}\in\Theta_{\boldsymbol{\Sigma}}(r)} \|\Delta(\boldsymbol{\beta})\|_2$ in (A.26), which only depends on the data $\mathscr{D}_n = \{(y_i, \boldsymbol{x}_i)\}_{i=1}^n$, has been dealt with in the proof of Theorem 2.1.2. Hence, it remains to bound the random fluctuation $\Delta^\flat(\boldsymbol{\beta}) - \mathbb{E}^*\Delta^\flat(\boldsymbol{\beta}) = \mathbf{S}^{-1/2}\{R_n^\flat(\boldsymbol{\beta}) - R_n^\flat(\boldsymbol{\beta}^*)\}$ over $\boldsymbol{\beta} \in \Theta_{\boldsymbol{\Sigma}}(r)$, given $\mathscr{D}_n$. As before, we use a change of variable $\boldsymbol{v} = \boldsymbol{\Sigma}^{1/2}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)$ and obtain

$$\begin{aligned}
&\sup_{\boldsymbol{\beta}\in\Theta_{\boldsymbol{\Sigma}}(r)} \|\Delta^\flat(\boldsymbol{\beta}) - \mathbb{E}^*\Delta^\flat(\boldsymbol{\beta})\|_2 = \sup_{\boldsymbol{\beta}\in\Theta_{\boldsymbol{\Sigma}}(r)} \|\mathbf{S}^{-1/2}\{R_n^\flat(\boldsymbol{\beta}) - R_n^\flat(\boldsymbol{\beta}^*)\}\|_2 \\
&\le \underline{f}^{-1/2} \sup_{\boldsymbol{\beta}\in\Theta_{\boldsymbol{\Sigma}}(r), \boldsymbol{u}\in\mathbb{B}^p(1)} \langle R_n^\flat(\boldsymbol{\beta}) - R_n^\flat(\boldsymbol{\beta}^*), \boldsymbol{\Sigma}^{-1/2}\boldsymbol{u}\rangle \\
&= \underline{f}^{-1/2} r^{-1} \sup_{\boldsymbol{u},\boldsymbol{v}\in\mathbb{B}^p(r)} \underbrace{\langle \boldsymbol{\Sigma}^{-1/2}\{R_n^\flat(\boldsymbol{\beta}^* + \boldsymbol{\Sigma}^{-1/2}\boldsymbol{v}) - R_n^\flat(\boldsymbol{\beta}^*)\}, \boldsymbol{v}\rangle}_{n^{-1/2}\Delta_0^\flat(\boldsymbol{u},\boldsymbol{v})},
\end{aligned} \tag{A.27}$$

where $\Delta_0^\flat(\boldsymbol{u}, \boldsymbol{v}) = n^{-1/2}\sum_{i=1}^n e_i\langle \boldsymbol{z}_i, \boldsymbol{u}\rangle\{I(\varepsilon_i \le \langle \boldsymbol{z}_i, \boldsymbol{v}\rangle) - I(\varepsilon_i \le 0)\}$. Let $\mathscr{F}_1$ and $\mathscr{F}_2$ be the function classes defined in (A.16), and let $\mathscr{F} = \mathscr{F}_1 \cdot (\mathscr{F}_2 - f_0)$ be the pointwise product between $\mathscr{F}_1$ and $\mathscr{F}_2 - f_0$ with $f_0 : (z_0, \boldsymbol{z}) \mapsto I(z_0 \le 0)$. With this notation, we have $\sup_{\boldsymbol{u},\boldsymbol{v}\in\mathbb{B}^p(r)} \Delta_0^\flat(\boldsymbol{u}, \boldsymbol{v}) = \sup_{f\in\mathscr{F}} n^{-1/2}\sum_{i=1}^n e_i f(\bar{\boldsymbol{z}}_i)$. Recall that $\mathbb{E}^*$ denotes the conditional expectation given $\mathscr{D}_n$. By Theorem 13 in Boucheron et al. [2005] and the bound $\sup_{1\le i\le n, f\in\mathscr{F}} f(\bar{\boldsymbol{z}}_i) \le r\max_{1\le i\le n}\|\boldsymbol{z}_i\|_2$, we obtain that, with $Z := \mathbb{E}^*\{\sup_{f\in\mathscr{F}} |(1/n)\sum_{i=1}^n e_i f(\bar{\boldsymbol{z}}_i)|\}$ denoting the conditional Rademacher

average,

$$\big\{\mathbb{E}(Z-\mathbb{E}Z)_+^{2k}\big\}^{1/(2k)} \le 2\sqrt{\mathbb{E}Z\cdot k\kappa r\frac{M_{n,k}}{n}} + 2k\kappa r\frac{M_{n,k}}{n} \le \mathbb{E}Z + 3k\kappa r\frac{M_{n,k}}{n}, \text{ valid for any } k\ge 1,$$

where $\kappa = \sqrt{e}/(2\sqrt{e}-2) < 1.271$ and $M_{n,k} := (\mathbb{E}\max_{1\le i\le n}\|\boldsymbol{z}_i\|_2^{2k})^{1/(2k)}$. By (A.27), Markov's inequality and the bound $Z \le (Z-\mathbb{E}Z)_+ + \mathbb{E}Z$, we obtain that

$$\sup_{\boldsymbol{\beta}\in\Theta_{\boldsymbol{\Sigma}}(r)} \|\Delta^\flat(\boldsymbol{\beta}) - \mathbb{E}^*\Delta^\flat(\boldsymbol{\beta})\|_2 = O_{\mathbb{P}^*}\big(r^{-1}Z\big) \text{ and } Z = O_{\mathbb{P}}\big(\mathbb{E}Z + rM_{n,1}/n\big). \tag{A.28}$$

For $\mathbb{E}Z$, by a similar argument to (A.20) and (A.21), we get

$$\mathbb{E}Z \lesssim r^{3/2}\sqrt{\log(C_2p/r)\frac{p}{n}} + r(p+\log n)^{1/2}\log(C_2p/r)\frac{p}{n}. \tag{A.29}$$

With the above preparations, we are ready to prove the claim. Together, Theorems 2.1.1–2.1.4 imply that under the scaling $n \gtrsim p+\log n$, there exists some event $\mathscr{E}_n$, satisfying $\mathbb{P}(\mathscr{E}_n) \ge 1-4n^{-1}$, on which $\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma}} \le r_n = C_3\sqrt{(p+\log n)/n}$ and

$$\begin{aligned}\chi_{1n} &:= \bigg\|\mathbf{S}^{1/2}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*) + \mathbf{S}^{-1/2}\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\big\{I(\varepsilon_i\le 0)-\tau\big\}\bigg\|_2 \\ &\le \sup_{\boldsymbol{\beta}\in\Theta_{\boldsymbol{\Sigma}}(r_n)}\|\Delta(\boldsymbol{\beta})\|_2 \lesssim \underbrace{\frac{(p+\log n)^{1/4}(p\log n)^{1/2}}{n^{3/4}} + (p+\log n)^{1/2}\frac{p\log n}{n}}_{:=\Delta_{n,p}}.\end{aligned}$$

Moreover, with $\mathbb{P}^*$-probability at least $1-n^{-1}$ conditioned on $\mathscr{E}_n$, $\|\widehat{\boldsymbol{\beta}}^\flat - \boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma}} \le r_n$ so that

$$\bigg\|\mathbf{S}^{1/2}(\widehat{\boldsymbol{\beta}}^\flat-\boldsymbol{\beta}^*) + \mathbf{S}^{-1/2}\frac{1}{n}\sum_{i=1}^n w_i\boldsymbol{x}_i\big\{I(\varepsilon_i\le 0)-\tau\big\}\bigg\|_2 \le \sup_{\boldsymbol{\beta}\in\Theta_{\boldsymbol{\Sigma}}(r_n)}\|\Delta^\flat(\boldsymbol{\beta})\|_2.$$

By (A.26), (A.28), (A.29) and (A.21), $\chi_{2n} = \chi_{2n}(\mathscr{D}_n) := \mathbb{E}^*\{\sup_{\boldsymbol{\beta}\in\Theta_{\boldsymbol{\Sigma}}(r_n^\flat)}\|\Delta^\flat(\boldsymbol{\beta})\|_2\}$ satisfies $\chi_{2n} = O_{\mathbb{P}}(\Delta_{n,d})$. Let $\boldsymbol{r}_n^\flat = \mathbf{S}^{1/2}(\widehat{\boldsymbol{\beta}}^\flat - \widehat{\boldsymbol{\beta}}) - \mathbf{S}^{-1/2}(1/n)\sum_{i=1}^n e_i\boldsymbol{x}_i\{\tau - I(\varepsilon_i\le 0)\}$. Then, with

$\mathbb{P}^*$-probability at least $1-n^{-1}$ conditioned on $\mathscr{E}_n$, $\|\boldsymbol{r}_n^\flat\|_2 \le \chi_{1n} + \sup_{\boldsymbol{\beta}\in\Theta_{\boldsymbol{\Sigma}}(r)} \|\Delta^\flat(\boldsymbol{\beta})\|_2$ with $\sup_{\boldsymbol{\beta}\in\Theta_{\boldsymbol{\Sigma}}(r)} \|\Delta^\flat(\boldsymbol{\beta})\|_2 = O_{\mathbb{P}^*}(\chi_{2n})$ and $\chi_{1n}+\chi_{2n} = O_{\mathbb{P}}(\Delta_{n,p})$. This establishes the claim result (2.14). □

A.1.7 Proof of Theorem 2.1.6

Let $\boldsymbol{\lambda} \in \mathbb{R}^p$ be an arbitrary vector defining a linear contrast of interest. Write $\gamma_i = \langle \mathbf{S}^{-1}\boldsymbol{\lambda}, \boldsymbol{x}_i\rangle$ and $\zeta_i = I(\varepsilon_i \le 0) - \tau$ for $i = 1,\ldots,n$, and define

$$S_n = \frac{1}{\sqrt{n}}\sum_{i=1}^n \gamma_i\zeta_i \quad \text{and} \quad S_n^\flat = \frac{1}{\sqrt{n}}\sum_{i=1}^n e_i\gamma_i\zeta_i.$$

It follows from Theorem 2.1.2 that under the scaling $n \gtrsim p+\log n$, there exists a sequence of events $\{\mathscr{E}_n\}$ with $\mathbb{P}(\mathscr{E}_n) \ge 1-4n^{-1}$ such that, $|n^{1/2}\langle \boldsymbol{\lambda}, \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\rangle - S_n| \le c_1\|\mathbf{S}^{-1/2}\boldsymbol{\lambda}\|_2\delta_{n,d}$ on $\mathscr{E}_n$, where $\delta_{n,p} := (p+\log n)^{1/4}(p\log n)^{1/2}n^{-1/4} + (p+\log n)^{1/2}p\log(n)n^{-1/2}$. By Theorems 2.1.4 and 2.1.5, we further have $|n^{1/2}\langle \boldsymbol{\lambda}, \widehat{\boldsymbol{\beta}}^\flat - \widehat{\boldsymbol{\beta}}\rangle - S_n^\flat| \le \|\mathbf{S}^{-1/2}\boldsymbol{\lambda}\|_2\|n^{1/2}\boldsymbol{r}_n^\flat\|_2$ with $\mathbb{P}^*$-probability at least $1-n^{-1}$ conditioned on $\mathscr{E}_n$. For the remainder $\boldsymbol{r}_n^\flat = \boldsymbol{r}_n^\flat(\{(e_i,y_i,\boldsymbol{x}_i)\}_{i=1}^n)$, using Markov's inequality with the bounds (A.28) and (A.29), there exits some event $\mathscr{G}_n$ with $\mathbb{P}(\mathscr{G}_n^{\mathrm{c}}) \lesssim (\delta_{n,d}/\delta_2)^2$ such that, conditioned on $\mathscr{E}_n \cap \mathscr{G}_n$,

$$\mathbb{P}^*\big(\|n^{1/2}\boldsymbol{r}_n^\flat\|_2 \ge \delta_1\big) \lesssim \delta_1^{-1}(\delta_{n,p}+\delta_2),$$

valid for any $\delta_1,\delta_2 > 0$. Taking $\delta_1 = \delta_{n,p}^{2/5}$ and $\delta_2 = \delta_{n,p}^{4/5}$ yields that $\mathbb{P}(\mathscr{G}_n^{\mathrm{c}}) \le c_2\delta_{n,p}^{2/5}$ and

$$\mathbb{P}^*\big(\|n^{1/2}\boldsymbol{r}_n^\flat\|_2 \ge \delta_{n,p}^{2/5}\big) \le c_3\delta_{n,p}^{2/5}, \text{ conditioned on } \mathscr{E}_n\cap\mathscr{G}_n.$$

Next we establish the closeness in distribution between S_n and $S_n^\flat$. Note that $\gamma_i\zeta_i$ are independent random variables with mean zero and $\mathrm{var}(\gamma_i\zeta_i) = \tau(1-\tau)\|\mathbf{S}^{-1}\boldsymbol{\lambda}\|_{\boldsymbol{\Sigma}}^2$. Thus,

$\mathrm{var}(S_n) = \tau(1-\tau)\|\mathbf{S}^{-1}\boldsymbol{\lambda}\|_{\boldsymbol{\Sigma}}^2 \geq \tau(1-\tau)\bar{f}^{-1}\|\mathbf{S}^{-1/2}\boldsymbol{\lambda}\|_2^2$. Moreover, under Condition 2.1.1,

$$\mathbb{E}\big(|\gamma_i\zeta_i|^3\big) \leq \tau(1-\tau)\mathbb{E}|\langle \mathbf{S}^{-1}\boldsymbol{\lambda}, \boldsymbol{x}_i\rangle|^3 \leq \tau(1-\tau)m_3\|\mathbf{S}^{-1}\boldsymbol{\lambda}\|_{\boldsymbol{\Sigma}}^3.$$

Let $\Phi(\cdot)$ be the standard normal distribution function. By the Berry-Esseen inequality (see, e.g., Tyurin [2011]),

$$\sup_{x\in\mathbb{R}}\big|\mathbb{P}\big\{S_n \leq \mathrm{var}(S_n)^{1/2}x\big\} - \Phi(x)\big| \leq \frac{m_3}{2\sqrt{\tau(1-\tau)n}}. \tag{A.30}$$

For $S_n^\flat$, using a conditional version of the Berry-Esseen inequality for sums of independent random variables [Tyurin, 2011], we have

$$\sup_{x\in\mathbb{R}}\big|\mathbb{P}^*\big\{S_n^\flat \leq \mathrm{var}^*(S_n^\flat)^{1/2}x\big\} - \Phi(x)\big| \leq \frac{(1/n)\sum_{i=1}^n|\zeta_i\gamma_i|^3}{2\sqrt{n}\{\mathrm{var}^*(S_n^\flat)\}^{3/2}}, \tag{A.31}$$

where $\mathrm{var}^*(S_n^\flat) = (1/n)\sum_{i=1}^n(\gamma_i\zeta_i)^2$. Recall that $\boldsymbol{z}_i = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}_i$, and let $\boldsymbol{u} = \boldsymbol{\Sigma}^{1/2}\mathbf{S}^{-1}\boldsymbol{\lambda}/\|\mathbf{S}^{-1}\boldsymbol{\lambda}\|_{\boldsymbol{\Sigma}} \in \mathbb{S}^{p-1}$ be a unit vector. For the two data-dependent quantities $\mathrm{var}^*(S_n^\flat)$ and $(1/n)\sum_{i=1}^n|\gamma_i\zeta_i|^3$, we have

$$\big|\mathrm{var}^*(S_n^\flat)/\mathrm{var}(S_n) - 1\big| = \frac{1}{\tau(1-\tau)}\left|\frac{1}{n}\sum_{i=1}^n\zeta_i^2\langle\boldsymbol{u},\boldsymbol{z}_i\rangle^2 - \tau(1-\tau)\right| \tag{A.32}$$

and

$$\frac{1}{n}\sum_{i=1}^n|\gamma_i\zeta_i|^3 \leq \max_{1\leq i\leq n}|\gamma_i\zeta_i|\cdot\frac{1}{n}\sum_{i=1}^n\zeta_i^2\langle\mathbf{S}^{-1}\boldsymbol{\lambda},\boldsymbol{x}_i\rangle^2 \leq \max_{1\leq i\leq n}|\gamma_i\zeta_i|\cdot\|\mathbf{S}^{-1}\boldsymbol{\lambda}\|_{\boldsymbol{\Sigma}}^2\cdot\frac{1}{n}\sum_{i=1}^n\zeta_i^2\langle\boldsymbol{u},\boldsymbol{z}_i\rangle^2. \tag{A.33}$$

For independent zero-mean sub-Gaussian random variables $\gamma_i\zeta_i$, it can be shown that with probability at least $1-e^{-x}$, $\max_{1\leq i\leq n}|\gamma_i\zeta_i| \lesssim \|\mathbf{S}^{-1}\boldsymbol{\lambda}\|_{\boldsymbol{\Sigma}}\sqrt{\log(n)+x}$. Furthermore, following

the proof of Proposition A.1.3, it can be similarly shown that

$$\left|\frac{1}{n}\sum_{i=1}^{n}\zeta_i^2\langle \boldsymbol{u}, \boldsymbol{z}_i\rangle^2 - \tau(1-\tau)\right| \le 2\upsilon_0^2\sqrt{\frac{2x}{3n}} + 2\upsilon_0^2\frac{x}{n}$$

with probability at least $1-2e^{-x}$. Putting together the pieces, it follows from (A.32) that there exists an event $\mathscr{E}_n'$, satisfying $\mathbb{P}(\mathscr{E}_n') \ge 1-n^{-1}$, on which $\max_{1\le i\le n}|\gamma_i\zeta_i| \lesssim \|\mathbf{S}^{-1}\boldsymbol{\lambda}\|_{\boldsymbol{\Sigma}}(\log n)^{1/2}$,

$$\frac{1}{n}\sum_{i=1}^{n}|\gamma_i\zeta_i|^3 \lesssim \|\mathbf{S}^{-1}\boldsymbol{\lambda}\|_{\boldsymbol{\Sigma}}^3(\log n)^{1/2} \text{ and } \left|\mathrm{var}^*(S_n^\flat)/\mathrm{var}(S_n) - 1\right| \lesssim \sqrt{\frac{\log n}{n}} \tag{A.34}$$

as long as $n \gtrsim \log n$.

For the normal distribution function, we have the following property derived from Pinsker's inequality (see Lemma A.7 in the supplement of Spokoiny and Zhilova [2015]):

$$\sup_{x\in\mathbb{R}}\left|\Phi(x/\mathrm{var}(S_n)^{1/2}) - \Phi(x/\mathrm{var}^*(S_n^\flat)^{1/2})\right| \le \frac{1}{2}\left|\mathrm{var}^*(S_n^\flat)/\mathrm{var}(S_n) - 1\right| \tag{A.35}$$

as long as $|\mathrm{var}^*(S_n^\flat)/\mathrm{var}(S_n) - 1| \le 1/2$. Moreover, for any $a \le b$,

$$\Phi(b/\mathrm{var}(S_n)^{1/2}) - \Phi(a/\mathrm{var}(S_n)^{1/2}) \le \frac{b-a}{\sqrt{2\pi\mathrm{var}(S_n)}} \le \frac{\bar{f}^{1/2}(b-a)}{\|\mathbf{S}^{-1/2}\boldsymbol{\lambda}\|_2\sqrt{2\pi\tau(1-\tau)}}. \tag{A.36}$$

Combining the ingredients, we derive that for any $x \in \mathbb{R}$,

$$
\begin{aligned}
&\mathbb{P}\big(n^{1/2}\langle \boldsymbol{\lambda}, \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \rangle \le x\big) \le \mathbb{P}\big(S_n \le x + c_1 \|\mathbf{S}^{-1/2}\boldsymbol{\lambda}\|_2 \delta_{n,p}\big) + 4n^{-1} \\
&\overset{\text{(i)}}{\le} \mathbb{P}\big\{\mathrm{var}(S_n)^{1/2} G \le x + c_1 \|\mathbf{S}^{-1/2}\boldsymbol{\lambda}\|_2 \delta_{n,p}\big\} + \frac{m_3}{2\sqrt{\tau(1-\tau)n}} + 4n^{-1} \\
&\overset{\text{(ii)}}{\le} \mathbb{P}\big\{\mathrm{var}(S_n)^{1/2} G \le x - \|\mathbf{S}^{-1/2}\boldsymbol{\lambda}\|_2 \delta_{n,p}^{2/5}\big\} + \bar{f}^{1/2}\frac{c_1\delta_{n,p} + \delta_{n,p}^{2/5}}{\sqrt{2\pi\tau(1-\tau)}} + \frac{m_3}{2\sqrt{\tau(1-\tau)n}} + 4n^{-1} \\
&\overset{\text{(iii)}}{\le} \mathbb{P}^*\big\{\mathrm{var}^*(S_n^\flat)^{1/2} G \le x - \|\mathbf{S}^{-1/2}\boldsymbol{\lambda}\|_2 \delta_{n,p}^{2/5}\big\} \\
&\quad + \frac{1}{2}\left|\frac{\mathrm{var}^*(S_n^\flat)}{\mathrm{var}(S_n)} - 1\right| + \bar{f}^{1/2}\frac{c_1\delta_{n,p} + \delta_{n,p}^{2/5}}{\sqrt{2\pi\tau(1-\tau)}} + \frac{m_3}{2\sqrt{\tau(1-\tau)n}} + 4n^{-1} \\
&\overset{\text{(iv)}}{\le} \mathbb{P}^*\big(S_n^\flat \le x - \|\mathbf{S}^{-1/2}\boldsymbol{\lambda}\|_2 \delta_{n,p}^{2/5}\big) + \frac{(1/n)\sum_{i=1}^n |\gamma_i\zeta_i|^3}{2\sqrt{n}\,\{\mathrm{var}^*(S_n^\flat)\}^{3/2}} \\
&\quad + \frac{1}{2}\left|\frac{\mathrm{var}^*(S_n^\flat)}{\mathrm{var}(S_n)} - 1\right| + \bar{f}^{1/2}\frac{c_1\delta_{n,p} + \delta_{n,p}^{2/5}}{\sqrt{2\pi\tau(1-\tau)}} + \frac{m_3}{2\sqrt{\tau(1-\tau)n}} + 4n^{-1},
\end{aligned}
$$

where steps (i) and (iv) follow respectively from the Berry-Esseen inequalities (A.30) and (A.31), step (ii) uses the anti-concentration inequality (A.36), and step (iii) is due to the Gaussian comparison inequality (A.35). Conditioned on $\mathscr{E}_n \cap \mathscr{G}_n$,

$$
\begin{aligned}
&\mathbb{P}^*\big(S_n^\flat \le x - \|\mathbf{S}^{-1/2}\boldsymbol{\lambda}\|_2 \delta_{n,p}^{2/5}\big) \\
&\le \mathbb{P}^*\big(S_n^\flat \le x - \|\mathbf{S}^{-1/2}\boldsymbol{\lambda}\|_2 \|n^{1/2}\boldsymbol{r}_n^\flat\|_2\big) + \mathbb{P}^*\big(\|n^{1/2}\boldsymbol{r}_n^\flat\|_2 \ge \delta_{n,p}^{2/5}\big) \\
&\le \mathbb{P}^*\big(n^{1/2}\langle \boldsymbol{\lambda}, \widehat{\boldsymbol{\beta}}^\flat - \widehat{\boldsymbol{\beta}} \rangle \le x\big) + n^{-1} + c_3\delta_{n,p}^{2/5}.
\end{aligned}
$$

Moreover, on the event $\mathscr{E}_n'$, the bounds in (A.34) imply

$$
\frac{(1/n)\sum_{i=1}^n |\gamma_i\zeta_i|^3}{2\sqrt{n}\,\{\mathrm{var}^*(S_n^\flat)\}^{3/2}} + \frac{1}{2}\left|\frac{\mathrm{var}^*(S_n^\flat)}{\mathrm{var}(S_n)} - 1\right| \lesssim \sqrt{\frac{\log n}{n}}
$$

as long as $n \gtrsim \log n$. A similar argument leads to a series of reverse inequalities.

Putting together the pieces, we conclude that conditioned on the event $\mathscr{E}_n \cap \mathscr{E}_n' \cap \mathscr{G}_n$,

$$\sup_{x \in \mathbb{R}} \left| \mathbb{P}\big(n^{1/2} \langle \boldsymbol{\lambda}, \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \rangle \leq x\big) - \mathbb{P}^*\big(n^{1/2} \langle \boldsymbol{\lambda}, \widehat{\boldsymbol{\beta}}^\flat - \widehat{\boldsymbol{\beta}} \rangle \leq x\big) \right| \lesssim \delta_{n,p}^{2/5}.$$

Under the scaling $p^3 (\log n)^2 = o(n)$, $\delta_{n,p} = o(1)$ as $n \to \infty$. Combined with the above bound, this establishes the claim (2.15). $\square$

Appendix B

Supplementary Material for Chapter 3

B.1 One-step Conquer with Higher-order Kernels

As noted in Section 3.3.1, the smoothing bias is of order h^2 when a non-negative kernel is used. The ensuing empirical loss $\boldsymbol{\beta} \mapsto (1/n)\sum_{i=1}^n (\rho_\tau * K_h)(y_i - \langle \boldsymbol{x}_i, \boldsymbol{\beta}\rangle)$ is not only twice-differentiable and convex, but also (provably) strongly convex in a local vicinity of $\boldsymbol{\beta}^*$ with high probability. Kernel smoothing is ubiquitous in nonparametric statistics. The order of a kernel, ν, is defined as the order of the first non-zero moment. The order of a symmetric kernel is always even. A kernel is called *high-order* if $\nu > 2$, which inevitably has negative parts and thus is no longer a probability density. Thus far we have focused on conquer with second-order kernels, and the resulting estimator achieves an ℓ_2-error of the order $\sqrt{p/n} + h^2$.

Let $G(\cdot)$ be a higher-order symmetric kernel with order $\nu \geq 4$, and $b > 0$ be a bandwidth. Again, via convolution smoothing, we may consider a bias-reduced estimator that minimizes the empirical loss $\boldsymbol{\beta} \mapsto \widehat{Q}_b^G(\boldsymbol{\beta}) := (1/n)\sum_{i=1}^n (\rho_\tau * G_b)(y_i - \langle \boldsymbol{x}_i, \boldsymbol{\beta}\rangle)$. This, however, leads to a non-convex optimization. Without further assumptions, finding a global minimum is computationally intractable: finding an ε-suboptimal point for a k-times continuously differentiable loss function requires at least $\Omega\{(1/\varepsilon)^{p/k}\}$ evaluations of the function and its first k derivatives, ignoring problem-dependent constants; see Section 1.6 in Nemirovski and Yudin [1983]. Instead, various gradient-based methods have been developed for computing *stationary points*, which are points $\boldsymbol{\beta}$ with sufficiently small gradient $\|\nabla \widehat{Q}_b^G(\boldsymbol{\beta})\|_2 \leq \varepsilon$, where $\varepsilon \geq 0$ is optimization error. However, the

equation $\nabla \widehat{Q}_b^G(\boldsymbol{\beta}) = \mathbf{0}$ does not necessarily have a unique solution, whose statistical guarantees remain unknown.

Motivated by the classical one-step estimator [Bickel, 1975], we further propose a one-step conquer estimator using high-order kernels, which bypasses solving a large-scale non-convex optimization. To begin with, we choose two symmetric kernel functions, $K : \mathbb{R} \mapsto [0, \infty)$ with order two and $G(\cdot)$ with order $\nu \geq 4$, and let $h, b > 0$ be two bandwidths. First, compute an initial conquer estimator $\overline{\boldsymbol{\beta}} \in \operatorname{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^p} \widehat{Q}_h^K(\boldsymbol{\beta})$, where $\widehat{Q}_h^K(\boldsymbol{\beta}) = (1/n)\sum_{i=1}^n (\rho_\tau * K_h)(y_i - \langle \boldsymbol{x}_i, \boldsymbol{\beta} \rangle)$. Denote by $\bar{\varepsilon}_i = y_i - \langle \boldsymbol{x}_i, \overline{\boldsymbol{\beta}} \rangle$ for $i = 1, \ldots, n$ the fitted residuals. Next, with slight abuse of notation, we define the one-step conquer estimator $\widehat{\boldsymbol{\beta}}$ as a solution to the equation $\nabla^2 \widehat{Q}_b^G(\overline{\boldsymbol{\beta}})(\widehat{\boldsymbol{\beta}} - \overline{\boldsymbol{\beta}}) = -\nabla \widehat{Q}_b^G(\overline{\boldsymbol{\beta}})$, or equivalently,

$$\left\{ \frac{1}{n} \sum_{i=1}^n G_b(\bar{\varepsilon}_i) \boldsymbol{x}_i \boldsymbol{x}_i^{\mathrm{T}} \right\} (\widehat{\boldsymbol{\beta}} - \overline{\boldsymbol{\beta}}) = \frac{1}{n} \sum_{i=1}^n \big\{ \mathscr{G}_b(\bar{\varepsilon}_i) + \tau - 1 \big\} \boldsymbol{x}_i. \tag{B.1}$$

where $\widehat{Q}_b^G(\boldsymbol{\beta}) = (1/n)\sum_{i=1}^n (\rho_\tau * G_b)(y_i - \langle \boldsymbol{x}_i, \boldsymbol{\beta} \rangle)$ and $\mathscr{G}_b(u) = \int_{-\infty}^{u/b} G(v)\,\mathrm{d}v$. Provided that $\nabla^2 \widehat{Q}_b^G(\overline{\boldsymbol{\beta}})$ is positive definite, the one-step conquer estimate $\widehat{\boldsymbol{\beta}}$ essentially performs a Newton-type step based on $\overline{\boldsymbol{\beta}}$:

$$\widehat{\boldsymbol{\beta}} = \overline{\boldsymbol{\beta}} - \big\{ \nabla^2 \widehat{Q}_b^G(\overline{\boldsymbol{\beta}}) \big\}^{-1} \nabla \widehat{Q}_b^G(\overline{\boldsymbol{\beta}}). \tag{B.2}$$

In general, $\widehat{\boldsymbol{\beta}}$ can be computed by the conjugate gradient method [Hestenes and Stiefel, 1952].

Theoretical properties of the one-step estimator $\widehat{\boldsymbol{\beta}}$ defined in (B.1), including the Bahadur representation and asymptotic normality with explicit Berry-Esseen bound, will be provided in on-line supplementary materials. For practical implementation, we consider higher-order Gaussian-based kernels. For $r = 1, 2, \ldots$, the $(2r)$-th order Gaussian kernels are

$$G(u; 2r) = \frac{(-1)^r \phi^{(2r-1)}(u)}{2^{r-1}(r-1)!u} = \sum_{\ell=0}^{r-1} \frac{(-1)^\ell}{2^\ell \ell!} \phi^{(2\ell)}(u);$$

see Section 2 of Wand and Schucany [1990]. Integrating $G(\cdot;2r)$ yields

$$\mathscr{G}(v;2r) = \int_{-\infty}^{v} G(u;2r)\,\mathrm{d}u = \sum_{\ell=0}^{r-1} \frac{(-1)^\ell}{2^\ell \ell!} \phi^{(2\ell-1)}(v).$$

In fact, both $G(\cdot;2r)$ and $\mathscr{G}(\cdot;2r)$ have simpler forms $G(u;2r) = p_r(u)\phi(u)$ and $\mathscr{G}(u;2r) = \Phi(u) + P_r(u)\phi(u)$, where $p_r(\cdot)$ and $P_r(\cdot)$ are polynomials in u. For example, $p_1(u) = 1$, $P_1(u) = 0$, $p_2(u) = (-u^2+3)/2$, $P_2(u) = u/2$, $p_3(u) = (u^4 - 10u^2 + 15)/8$, and $P_3(u) = (-u^3 + 7u)/8$. We refer to Oryshchenko [2020] for more details when r is large.

B.2 Proofs for Section 3.3

Recall that $\boldsymbol{x} = (x_1, \ldots, x_p)^{\mathrm{T}}$ is such that $x_1 \equiv 1$, $\mathbb{E}(x_j) = 0$ for $j = 2, \ldots, p$, and $\boldsymbol{\Sigma} = \mathbb{E}(\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}})$ is positive definite. In this case,

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & \mathbf{0}_{p-1}^{\mathrm{T}} \\ \mathbf{0}_{p-1} & \mathbf{S} \end{bmatrix} \quad \text{with} \quad \mathbf{S} = \mathbb{E}(\boldsymbol{x}_{-}\boldsymbol{x}_{-}^{\mathrm{T}}) \quad \text{and} \quad \boldsymbol{w} = \begin{bmatrix} 1 \\ \mathbf{S}^{-1/2}\boldsymbol{x}_{-} \end{bmatrix},$$

where $\mathbf{0}_k$ is the zero vector in $\mathbb{R}^k$ ($k \geq 2$). For every $r \geq 0$, define the ellipse $\Theta(r) = \{\boldsymbol{u} \in \mathbb{R}^p : \|\boldsymbol{u}\|_{\boldsymbol{\Sigma}} \leq r\}$ and its boundary $\partial\Theta(r) = \{\boldsymbol{u} \in \mathbb{R}^p : \|\boldsymbol{u}\|_{\boldsymbol{\Sigma}} = r\}$.

B.2.1 Proof of Proposition 3.3.1

To begin with, define $\boldsymbol{\delta}_h = \boldsymbol{\beta}_h^* - \boldsymbol{\beta}^* \in \mathbb{R}^p$ and $\delta_h = \|\boldsymbol{\delta}_h\|_{\boldsymbol{\Sigma}}$. By the convexity of $\boldsymbol{\beta} \mapsto Q_h(\boldsymbol{\beta})$ and the first-order optimality condition $\nabla Q_h(\boldsymbol{\beta}_h^*) = \mathbf{0}$, we have

$$0 \leq \langle \nabla Q_h(\boldsymbol{\beta}_h^*) - \nabla Q_h(\boldsymbol{\beta}^*), \boldsymbol{\beta}_h^* - \boldsymbol{\beta}^* \rangle = \langle -\nabla Q_h(\boldsymbol{\beta}^*), \boldsymbol{\delta}_h \rangle \leq \|\boldsymbol{\Sigma}^{-1/2} \nabla Q_h(\boldsymbol{\beta}^*)\|_2 \cdot \|\boldsymbol{\delta}_h\|_{\boldsymbol{\Sigma}}, \tag{B.3}$$

where the last step follows from Hölder's inequality. Note that $\nabla Q_h(\boldsymbol{\beta}^*) = \mathbb{E}\{\mathscr{K}(-\varepsilon/h) - \tau\}\boldsymbol{x}$. By integration by parts and a Taylor series expansion,

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

from which it follows that $|\mathbb{E}\{\mathscr{K}(-\varepsilon/h)|\boldsymbol{x}\} - \tau| \le 0.5 l_0 \kappa_2 h^2$. Consequently,

$$\|\boldsymbol{\Sigma}^{-1/2}\nabla Q_h(\boldsymbol{\beta}^*)\|_2 = \sup_{\boldsymbol{u}\in\mathbb{S}^{p-1}} \mathbb{E}\big\{\mathscr{K}(-\varepsilon/h) - \tau\big\}\langle \boldsymbol{u}, \boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}\rangle \le 0.5 l_0 \kappa_2 h^2. \tag{B.4}$$

Turning to the left-hand side of (B.3), applying the mean value theorem for vector-valued functions implies

$$\nabla Q_h(\boldsymbol{\beta}_h^*) - \nabla Q_h(\boldsymbol{\beta}^*) = \int_0^1 \nabla^2 Q_h(\boldsymbol{\beta}^* + t\boldsymbol{\delta}_h)\,\mathrm{d}t\,\boldsymbol{\delta}_h, \tag{B.5}$$

where $\nabla^2 Q_h(\boldsymbol{\beta}) = \mathbb{E}\big\{K_h(y - \langle \boldsymbol{x}, \boldsymbol{\beta}\rangle)\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\big\}$ for $\boldsymbol{\beta} \in \mathbb{R}^p$. With $\boldsymbol{\delta} = \boldsymbol{\beta} - \boldsymbol{\beta}^*$, note that

$$\mathbb{E}\big\{K_h(y - \langle \boldsymbol{x}, \boldsymbol{\beta}\rangle)|\boldsymbol{x}\big\} = \frac{1}{h}\int_{-\infty}^{\infty} K\left(\frac{u - \langle \boldsymbol{x}, \boldsymbol{\delta}\rangle}{h}\right) f_{\varepsilon|\boldsymbol{x}}(u)\,\mathrm{d}u = \int_{-\infty}^{\infty} K(v) f_{\varepsilon|\boldsymbol{x}}(\langle \boldsymbol{x}, \boldsymbol{\delta}\rangle + hv)\,\mathrm{d}v.$$

By the Lipschitz continuity of $f_{\varepsilon|\boldsymbol{x}}(\cdot)$,

$$\mathbb{E}\big\{K_h(y - \langle \boldsymbol{x}, \boldsymbol{\beta}\rangle)|\boldsymbol{x}\big\} = f_{\varepsilon|\boldsymbol{x}}(0) + R_h(\boldsymbol{\delta}) \tag{B.6}$$

with $R_h(\boldsymbol{\delta})$ satisfying $|R_h(\boldsymbol{\delta})| \le l_0\big(|\langle \boldsymbol{x}, \boldsymbol{\delta}\rangle| + \kappa_1 h\big)$. Together, (B.5), (B.6) and the assumption

$f_{\varepsilon|\boldsymbol{x}}(0) \geq \underline{f} > 0$ (almost surely) yield

$$\begin{aligned}
&\langle \nabla Q_h(\boldsymbol{\beta}_h^*) - \nabla Q_h(\boldsymbol{\beta}^*),\ \boldsymbol{\beta}_h^* - \boldsymbol{\beta}^* \rangle \\
&\geq \underline{f} \cdot \|\boldsymbol{\delta}_h\|_{\boldsymbol{\Sigma}}^2 - 0.5 l_0 \mathbb{E}|\langle \boldsymbol{x}, \boldsymbol{\delta}_h \rangle|^3 - l_0 \kappa_1 h \cdot \|\boldsymbol{\delta}_h\|_{\boldsymbol{\Sigma}}^2 \geq \underline{f} \cdot \delta_h^2 - 0.5 l_0 m_3 \cdot \delta_h^3 - l_0 \kappa_1 h \cdot \delta_h^2. \qquad \text{(B.7)}
\end{aligned}$$

Combining (B.3) with the upper and lower bounds (B.4) and (B.7), we find that $\delta_h \geq 0$ satisfies $0.5 l_0 m_3 \cdot \delta_h^2 - (\underline{f} - l_0 \kappa_1 h)\delta_h + 0.5 l_0 \kappa_2 h^2 \geq 0$. Provided that $l_0\{\kappa_1 + (m_3 \kappa_2)^{1/2}\}h > \underline{f}$, solving this inequality yields

$$\delta_h \leq \frac{l_0 \kappa_2 h^2}{\underline{f} - l_0 \kappa_1 h + \Delta_h^{1/2}} \quad \text{or} \quad \delta_h \geq \frac{\underline{f} - l_0 \kappa_1 h + \Delta_h^{1/2}}{l_0 m_3}. \qquad \text{(B.8)}$$

where $\Delta_h := (\underline{f} - l_0 \kappa_1 h)^2 - l_0^2 m_3 \kappa_2 h^2 > 0$. It remains to rule out the second bound in (B.8).

Assume δ_h satisfies the second bound in (B.8), so that $\delta_h > l_0 (m_3 \kappa_2)^{1/2} h / (l_0 m_3) = (\kappa_2 / m_3)^{1/2} h =: r_0$. Then, there exists some $\eta \in (0,1)$ such that $\widetilde{\boldsymbol{\beta}} := (1-\eta)\boldsymbol{\beta}^* + \eta \boldsymbol{\beta}_h^*$ satisfies $\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma}} = \eta \delta_h = r_0$. By the convexity of $\boldsymbol{\beta} \mapsto Q_h(\boldsymbol{\beta})$ and Lemma C.1 in the supplementary material of Sun, Zhou and Fan [2020],

$$\langle \nabla Q_h(\widetilde{\boldsymbol{\beta}}) - \nabla Q_h(\boldsymbol{\beta}^*), \widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \rangle \leq \eta \cdot \langle \nabla Q_h(\boldsymbol{\beta}_h^*) - \nabla Q_h(\boldsymbol{\beta}^*), \boldsymbol{\beta}_h^* - \boldsymbol{\beta}^* \rangle = \langle -\nabla Q_h(\boldsymbol{\beta}^*), \widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \rangle.$$

Repeating the above analysis, we find that the right-hand side of the above inequality $\leq 0.5 l_0 \kappa_2 h^2 \cdot r_0$, and the left-hand side

$$\begin{aligned}
\langle \nabla Q_h(\widetilde{\boldsymbol{\beta}}) - \nabla Q_h(\boldsymbol{\beta}^*), \widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \rangle &\geq \underline{f} \cdot r_0^2 - 0.5 l_0 m_3 \cdot r_0^3 - l_0 \kappa_1 h \cdot r_0^2 \\
&= \{\underline{f} - l_0 \kappa_1 h - 0.5 l_0 (m_3 \kappa_2)^{1/2} h\} r_0^2.
\end{aligned}$$

Canceling out the common factor r_0 from both sides, we obtain

$$r_0 \leq \frac{0.5 l_0 \kappa_2 h^2}{\underline{f} - l_0 \kappa_1 h - 0.5 l_0 (m_3 \kappa_2)^{1/2}} < \frac{0.5 l_0 \kappa_2 h^2}{0.5 l_0 (m_3 \kappa_2)^{1/2} h} = (\kappa_2 / m_3)^{1/2} h = r_0,$$

which leads to a contradiction. Consequently, δ_h must satisfy the first bound in (B.8), which in turn implies the claimed result.

Next, to investigate the leading term in the bias, define

$$\boldsymbol{\Delta} = \boldsymbol{\Sigma}^{-1/2} \big\{ \nabla Q_h(\boldsymbol{\beta}_h^*) - \nabla Q_h(\boldsymbol{\beta}^*) - \mathbf{J}(\boldsymbol{\beta}_h^* - \boldsymbol{\beta}^*) \big\} \quad \text{and} \quad \mathbf{H} = \boldsymbol{\Sigma}^{-1/2} \mathbf{J} \boldsymbol{\Sigma}^{-1/2} = \mathbb{E}\{ f_{\varepsilon|\boldsymbol{x}}(0) \boldsymbol{w} \boldsymbol{w}^{\mathrm{T}} \},$$

where $\boldsymbol{w} = \boldsymbol{\Sigma}^{-1/2} \boldsymbol{x}$. Again, by the mean value theorem for vector-valued functions,

$$\boldsymbol{\Delta} = \left\{ \boldsymbol{\Sigma}^{-1/2} \int_0^1 \nabla^2 Q_h(\boldsymbol{\beta}^* + t \boldsymbol{\delta}_h) \, \mathrm{d}t \, \boldsymbol{\Sigma}^{-1/2} - \mathbf{H} \right\} \boldsymbol{\Sigma}^{1/2} \boldsymbol{\delta}_h \quad \text{with} \quad \boldsymbol{\delta}_h = \boldsymbol{\beta}_h^* - \boldsymbol{\beta}^*. \tag{B.9}$$

The Lipschitz continuity of $f_{\varepsilon|\boldsymbol{x}}(\cdot)$ ensures that

$$\begin{aligned}
&\left\| \boldsymbol{\Sigma}^{-1/2} \int_0^1 \nabla^2 Q_h(\boldsymbol{\beta}^* + t \boldsymbol{\delta}_h) \, \mathrm{d}t \, \boldsymbol{\Sigma}^{-1/2} - \mathbf{H} \right\|_2 \\
&= \left\| \mathbb{E} \int_0^1 \int_{-\infty}^{\infty} K(u) \big\{ f_{\varepsilon|\boldsymbol{x}}(t \langle \boldsymbol{x}, \boldsymbol{\delta}_h \rangle - hu) - f_{\varepsilon|\boldsymbol{x}}(0) \big\} \, \mathrm{d}u \, \mathrm{d}t \, \boldsymbol{w} \boldsymbol{w}^{\mathrm{T}} \right\|_2 \\
&\leq l_0 \sup_{\boldsymbol{u} \in \mathbb{S}^{p-1}} \mathbb{E} \int_0^1 \int_{-\infty}^{\infty} K(u) \big(t |\langle \boldsymbol{x}, \boldsymbol{\delta}_h \rangle| + h|u| \big) \, \mathrm{d}u \, \mathrm{d}t \, \langle \boldsymbol{w}, \boldsymbol{u} \rangle^2 \\
&\leq 0.5 l_0 \sup_{\boldsymbol{u} \in \mathbb{S}^{p-1}} \mathbb{E} \big(|\langle \boldsymbol{x}, \boldsymbol{\delta}_h \rangle| \langle \boldsymbol{w}, \boldsymbol{u} \rangle^2 \big) + l_0 \kappa_1 h \\
&\leq 0.5 l_0 m_3 \| \boldsymbol{\delta}_h \|_{\boldsymbol{\Sigma}} + l_0 \kappa_1 h.
\end{aligned}$$

This bound, together with (B.9), implies

$$\| \boldsymbol{\Delta} \|_2 \leq l_0 \big(0.5 m_3 \| \boldsymbol{\delta}_h \|_{\boldsymbol{\Sigma}} + \kappa_1 h \big) \| \boldsymbol{\delta}_h \|_{\boldsymbol{\Sigma}}. \tag{B.10}$$

Moreover, applying a second-order Taylor series expansion to $f_{\varepsilon|\boldsymbol{x}}(\cdot)$ yields

$$\begin{aligned}
&\mathbb{E}\big\{\mathscr{K}(-\varepsilon/h)|\boldsymbol{x}\big\}-\tau \\
&=\int_{-\infty}^{\infty}K(u)\int_{0}^{-hu}\big\{f_{\varepsilon|\boldsymbol{x}}(t)-f_{\varepsilon|\boldsymbol{x}}(0)\big\}\,\mathrm{d}t\,\mathrm{d}u \\
&=0.5\kappa_2 h^2\cdot f'_{\varepsilon|\boldsymbol{x}}(0)+\int_{-\infty}^{\infty}\int_{0}^{-hu}\int_{0}^{t}K(u)\big\{f'_{\varepsilon|\boldsymbol{x}}(v)-f'_{\varepsilon|\boldsymbol{x}}(0)\big\}\,\mathrm{d}v\,\mathrm{d}t\,\mathrm{d}u.
\end{aligned}$$

For $\nabla Q_h(\boldsymbol{\beta}^*)=\mathbb{E}\{\mathscr{K}(-\varepsilon/h)-\tau\}\boldsymbol{x}$, it follows that

$$\left\|\boldsymbol{\Sigma}^{-1/2}\nabla Q_h(\boldsymbol{\beta}^*)-\frac{1}{2}\kappa_2 h^2\cdot\boldsymbol{\Sigma}^{-1/2}\mathbb{E}\big\{f'_{\varepsilon|\boldsymbol{x}}(0)\boldsymbol{x}\big\}\right\|_2\leq\frac{1}{6}l_1\kappa_3 h^3. \tag{B.11}$$

Combining (B.10) and (B.11) completes the proof of (3.21). □

B.2.2 Proof of Theorem 3.3.1

For every $\boldsymbol{\delta}\in\mathbb{R}^p$, define $\widehat{D}_h(\boldsymbol{\delta})=\widehat{Q}_h(\boldsymbol{\beta}^*+\boldsymbol{\delta})-\widehat{Q}_h(\boldsymbol{\beta}^*)$, $D_h(\boldsymbol{\delta})=Q_h(\boldsymbol{\beta}^*+\boldsymbol{\delta})-Q_h(\boldsymbol{\beta}^*)$, as well as first-order Taylor series remainder terms $\widehat{R}_h(\boldsymbol{\delta})=\widehat{D}_h(\boldsymbol{\delta})-\langle\nabla\widehat{Q}_h(\boldsymbol{\beta}^*),\boldsymbol{\delta}\rangle$ and $R_h(\boldsymbol{\delta})=D_h(\boldsymbol{\delta})-\langle\nabla Q_h(\boldsymbol{\beta}^*),\boldsymbol{\delta}\rangle$. With these notations, we have

$$\begin{aligned}
\widehat{D}_h(\boldsymbol{\delta})&=\langle\nabla Q_h(\boldsymbol{\beta}^*),\boldsymbol{\delta}\rangle+R_h(\boldsymbol{\delta})+\{\widehat{D}_h(\boldsymbol{\delta})-D_h(\boldsymbol{\delta})\} \\
&\geq R_h(\boldsymbol{\delta})-\|\boldsymbol{\Sigma}^{-1/2}\nabla Q_h(\boldsymbol{\beta}^*)\|_2\cdot\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}}-\{D_h(\boldsymbol{\delta})-\widehat{D}_h(\boldsymbol{\delta})\} \\
&\geq R_h(\boldsymbol{\delta})-0.5l_0\kappa_2 h^2\cdot\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}}-\underbrace{\{D_h(\boldsymbol{\delta})-\widehat{D}_h(\boldsymbol{\delta})\}}_{\text{sampling error}},
\end{aligned}\tag{B.12}$$

where we used (B.4) in the last step. Following the same argument that leads to (B.7), it can be shown that

$$R_h(\boldsymbol{\delta})\geq\frac{1}{2}\big(\underline{f}-l_0\kappa_1 h-0.5l_0 m_3\cdot\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}}\big)\cdot\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}}^2\ \text{ for all }\boldsymbol{\delta}\in\mathbb{R}^p. \tag{B.13}$$

Take $r_0 = (2\kappa_2/m_3)^{1/2}h$ as an intermediate convergence radius. For any $\boldsymbol{\delta} \in \partial\Theta(r_0)$, i.e., $\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}} = r_0$, the last two displays imply

$$R_h(\boldsymbol{\delta}) - 0.5 l_0 \kappa_2 h^2 \cdot \|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}} \geq \frac{1}{2}\big\{\underline{f} - l_0\kappa_1 h - l_0(2\kappa_2 m_3)^{1/2}h\big\} r_0^2 \text{ for all } \boldsymbol{\delta} \in \partial\Theta(r_0). \quad \text{(B.14)}$$

To control the last term on the right-hand side of (B.12), the following lemma provides some type of uniform law of large numbers for the zero-mean stochastic process $\{\widehat{D}_h(\boldsymbol{\delta}) - D_h(\boldsymbol{\delta}), \boldsymbol{\delta} \in \Theta(r)\}$, $r > 0$. This implies a form of restricted strong convexity (RSC) of the empirical loss $\widehat{Q}_h(\cdot)$.

Lemma B.2.1. *Given any $r \geq 0$, the bound*

$$\sup_{\boldsymbol{\delta}\in\Theta(r)} \{D_h(\boldsymbol{\delta}) - \widehat{D}_h(\boldsymbol{\delta})\} \leq 3\bar{\tau}\upsilon_0 r \cdot \left(\sqrt{\frac{u}{n}} + \frac{u}{n}\right) \quad \text{(B.15)}$$

holds with probability at least $1 - e^{4p-u}$ for any $u \geq 0$, where $\bar{\tau} = \max(\tau, 1-\tau)$. In addition, given any $r_u > r_l > 0$, with probability at least $1 - \lceil e\log(\frac{r_u}{r_l})\rceil e^{4p-u}$ for any $u \geq 0$,

$$D_h(\boldsymbol{\delta}) - \widehat{D}_h(\boldsymbol{\delta}) \leq 4.25\bar{\tau}\upsilon_0\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}}\left(\sqrt{\frac{u}{n}} + \frac{u}{n}\right) \text{ holds for all } \boldsymbol{\delta} \text{ satisfying } r_l \leq \|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}} \leq r_u. \quad \text{(B.16)}$$

First, applying (B.15) with $r = r_0$ and $u = 4p + t$ yields that, with probability at least $1 - e^{-t}$,

$$\sup_{\boldsymbol{\delta}\in\Theta(r_0)} \{D_h(\boldsymbol{\delta}) - \widehat{D}_h(\boldsymbol{\delta})\} \leq 3\bar{\tau}\upsilon_0 r_0 \left(\sqrt{\frac{4p+t}{n}} + \frac{4p+t}{n}\right). \quad \text{(B.17)}$$

Combining this bound with (B.12) and (B.14) implies that with probability at least $1 - e^{-t}$, $\widehat{D}_h(\boldsymbol{\delta}) > 0$ for all $\boldsymbol{\delta} \in \partial\Theta(r_0)$ as long as the bandwidth is subject to $\underline{f}^{-1} m_3^{1/2}\upsilon_0\sqrt{(p+t)/n} \lesssim h \lesssim \underline{f} m_3^{-1/2}$. On the other hand, by the optimality of $\widehat{\boldsymbol{\beta}}_h$, $\widehat{\boldsymbol{\delta}} := \widehat{\boldsymbol{\beta}}_h - \boldsymbol{\beta}^*$ satisfies $\widehat{D}_h(\widehat{\boldsymbol{\delta}}) \leq$

0. Consequently, the convexity of $\widehat{Q}_h(\cdot)$ ensures that $\|\widehat{\boldsymbol{\delta}}\|_{\boldsymbol{\Sigma}} \le r_0$. See, e.g., Lemma 9.21 in Wainwright [2019].

Next, we refine the convergence rate of $\widehat{\boldsymbol{\beta}}_h$ from r_0 to the claimed one under the above event. Consider the ring-shaped set $\Theta(r_l, r_0) = \{\boldsymbol{\delta} \in \mathbb{R}^p : r_l \le \|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}} \le r_0\}$ with $r_l = r_0 h$. If $\widehat{\boldsymbol{\delta}} \notin \Theta(r_l, r_0)$, we must have $\widehat{\boldsymbol{\delta}} \in \Theta(r_0 h)$, and thus the claimed bound follows immediately. Hereinafter, we assume $\widehat{\boldsymbol{\delta}} \in \Theta(r_l, r_0)$. Using the second inequality in Lemma B.2.1 with $(r_l, r_u) = (r_0 h, r_0)$ and $u = \sqrt{\log(e \log h^{-1}) + 4p + t}$, we find that with probability at least $1 - e^{-t}$,

$$D_h(\boldsymbol{\delta}) - \widehat{D}_h(\boldsymbol{\delta}) \le \|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}} \cdot \underbrace{4.25 \bar{\tau} \upsilon_0 \left\{ \sqrt{\frac{\log(e \log h^{-1}) + 4p + t}{n}} + \frac{\log(e \log h^{-1}) + 4p + t}{n} \right\}}_{=: r_1} \tag{B.18}$$

holds for all $\boldsymbol{\delta} \in \Theta(r_l, r_0)$, hence including $\widehat{\boldsymbol{\delta}}$. Applying this along with the earlier bounds (B.12), (B.13) and the fact $\widehat{D}_h(\widehat{\boldsymbol{\delta}}) \le 0$, we obtain that

$$(\underline{f} - l_0 \kappa_1 h) \|\widehat{\boldsymbol{\delta}}\|_{\boldsymbol{\Sigma}}^2 \le (2 r_1 + l_0 \kappa_2 h^2) \|\widehat{\boldsymbol{\delta}}\|_{\boldsymbol{\Sigma}} + 0.5 l_0 m_3 \|\widehat{\boldsymbol{\delta}}\|_{\boldsymbol{\Sigma}}^3 \le 2(r_1 + l_0 \kappa_2 h^2) \|\widehat{\boldsymbol{\delta}}\|_{\boldsymbol{\Sigma}}.$$

Canceling out $\|\widehat{\boldsymbol{\delta}}\|_{\boldsymbol{\Sigma}}$ proves the claimed bound. □

Proof of Lemma B.2.1

For each sample $\boldsymbol{z}_i = (\boldsymbol{x}_i, \varepsilon_i)$, define the loss difference $d_h(\boldsymbol{\delta}; \boldsymbol{z}_i) = \ell_h(\varepsilon_i - \langle \boldsymbol{x}_i, \boldsymbol{\delta} \rangle) - \ell_h(\varepsilon_i)$, so that $\widehat{D}_h(\boldsymbol{\delta}) = (1/n) \sum_{i=1}^n d_h(\boldsymbol{\delta}; \boldsymbol{z}_i)$. By the Lipschitz continuity of $u \mapsto \ell_h(u)$, $d_h(\boldsymbol{\delta}; \boldsymbol{z}_i)$ is $\bar{\tau}$-Lipschitz continuous in $\langle \boldsymbol{x}_i, \boldsymbol{\delta} \rangle$. That is, for any $\boldsymbol{z}_i$ and $\boldsymbol{\delta}, \boldsymbol{\delta}' \in \mathbb{R}^p$, $|d_h(\boldsymbol{\delta}; \boldsymbol{z}_i) - d_h(\boldsymbol{\delta}'; \boldsymbol{z}_i)| \le \bar{\tau} |\langle \boldsymbol{x}_i, \boldsymbol{\delta} \rangle - \langle \boldsymbol{x}_i, \boldsymbol{\delta}' \rangle|$.

For any given $r > 0$ and some $\varepsilon \in (0, 1)$ to be determined, define the random variable $\Delta_\varepsilon(r) = n(1 - \varepsilon) \sup_{\boldsymbol{\delta} \in \Theta(r)} \{D_h(\boldsymbol{\delta}) - \widehat{D}_h(\boldsymbol{\delta})\} / (2 \bar{\tau} r)$, where $D_h(\boldsymbol{\delta}) = \mathbb{E} \widehat{D}_h(\boldsymbol{\delta})$. By Chernoff's

inequality, for any $u \geq 0$,

$$\mathbb{P}\{\Delta_\varepsilon(r) \geq u\} \leq \exp\left[- \sup_{\lambda \geq 0} \left\{ \lambda u - \log \mathbb{E} e^{\lambda \Delta_\varepsilon(r)} \right\} \right]. \tag{B.19}$$

To control the moment generating function $\mathbb{E}e^{\lambda\Delta_\varepsilon(r)}$, by Rademacher symmetrization we have

$$\mathbb{E}e^{\lambda\Delta_\varepsilon(r)} \leq \mathbb{E}\exp\left\{ 2\lambda(1-\varepsilon) \sup_{\boldsymbol{\delta}\in\Theta(r)} \frac{1}{2\bar{\tau}r} \sum_{i=1}^n e_i d_h(\boldsymbol{\delta}; \boldsymbol{z}_i) \right\},$$

where $e_1, \ldots, e_n$ are independent Rademacher random variables. Recall that $d_h(\boldsymbol{\delta}; \boldsymbol{z}_i)$ is $\bar{\tau}$-Lipschitz continuous in $\langle \boldsymbol{x}_i, \boldsymbol{\delta} \rangle$, and $d_h(\boldsymbol{\delta}; \boldsymbol{z}_i) = 0$ if $\langle \boldsymbol{x}_i, \boldsymbol{\delta} \rangle = 0$. Applying the Ledoux-Talagrand contraction inequality (see Theorem 4.12 and inequality (4.10) in Ledoux and Talagrand [1991]) yields

$$\begin{aligned}
&\mathbb{E}\exp\left\{ 2\lambda(1-\varepsilon) \sup_{\boldsymbol{\delta}\in\Theta(r)} \frac{1}{2\bar{\tau}r} \sum_{i=1}^n e_i d_h(\boldsymbol{\delta}; \boldsymbol{z}_i) \right\} \\
&\leq \mathbb{E}\exp\left\{ \frac{\lambda}{r}(1-\varepsilon) \sup_{\boldsymbol{\delta}\in\Theta(r)} \sum_{i=1}^n e_i \langle \boldsymbol{x}_i, \boldsymbol{\delta} \rangle \right\} \leq \mathbb{E}\exp\left\{ \lambda(1-\varepsilon) \left\| \sum_{i=1}^n e_i \boldsymbol{w}_i \right\|_2 \right\},
\end{aligned}$$

where $\boldsymbol{w}_i = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}_i$. For this $\varepsilon \in (0,1)$, there exists an ε-net $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{N_\varepsilon}\}$ of $\mathbb{S}^{p-1}$ with cardinality $N_\varepsilon \leq (1+2/\varepsilon)^p$ such that $\|\sum_{i=1}^n e_i \boldsymbol{w}_i\|_2 \leq (1-\varepsilon)^{-1} \max_{1\leq j\leq N_\varepsilon} \sum_{i=1}^n e_i \boldsymbol{u}_j^{\mathrm{T}} \boldsymbol{w}_i$. This implies

$$\mathbb{E}\exp\left\{ \lambda(1-\varepsilon) \left\| \sum_{i=1}^n e_i \boldsymbol{w}_i \right\|_2 \right\} \leq \sum_{j=1}^{N_\varepsilon} \mathbb{E}\exp\left\{ \lambda \sum_{i=1}^n e_i \boldsymbol{u}_j^{\mathrm{T}} \boldsymbol{w}_i \right\}.$$

Write $S_j = \sum_{i=1}^n e_i \boldsymbol{u}_j^{\mathrm{T}} \boldsymbol{w}_i$, which is a sum of zero-mean random variables. Note that $e_i \in \{-1, 1\}$ is symmetric, and Condition 3.3.4 ensures that for any $k \geq 3$, $\mathbb{E}|\boldsymbol{u}_j^{\mathrm{T}} \boldsymbol{w}_i|^k \leq \upsilon_0^k k \int_0^\infty t^{k-1} e^{-t} \mathrm{d}t = \upsilon_0^k k!$.

Hence, for every $0 \le c < 1/\upsilon_0$,

$$\mathbb{E}e^{ce_i \boldsymbol{u}_j^{\mathrm{T}} \boldsymbol{w}_i} = 1 + \frac{c^2}{2}\mathbb{E}(e_i \boldsymbol{u}_j^{\mathrm{T}} \boldsymbol{w}_i)^2 + \sum_{\ell=3}^{\infty} \frac{c^\ell}{\ell!}\mathbb{E}(e_i \boldsymbol{u}_j^{\mathrm{T}} \boldsymbol{w}_i)^\ell$$
$$\le 1 + \frac{c^2}{2} + \sum_{\ell=2}^{\infty} \frac{c^{2\ell}}{(2\ell)!}\upsilon_0^{2\ell}(2\ell)! \le 1 + \frac{c^2}{2} + \sum_{\ell=2}^{\infty}(c^2\upsilon_0^2)^\ell \le 1 + \frac{\upsilon_0^2}{2}\sum_{\ell\ge 2} c^\ell(\sqrt{2}\upsilon_0)^{\ell-2}.$$

It follows that for every $0 < \lambda < 1/(\sqrt{2}\upsilon_0)$ and $j = 1, \ldots, N_\varepsilon$,

$$\log \mathbb{E}e^{\lambda S_j} \le \frac{n\upsilon_0^2\lambda^2}{2(1-\sqrt{2}\upsilon_0\lambda)} \text{ and thus } \log \mathbb{E}e^{\lambda \Delta_\varepsilon(r)} \le \log N_\varepsilon + \frac{n\upsilon_0^2\lambda^2}{2(1-\sqrt{2}\upsilon_0\lambda)}.$$

For any $u \ge 0$, note that

$$\sup_{\lambda \ge 0}\left\{\lambda u - \log \mathbb{E}e^{\lambda \Delta_\varepsilon(r)}\right\} \ge -\log N_\varepsilon + \sup_{\lambda \in (0,(\sqrt{2}\upsilon_0)^{-1})}\left\{\lambda u - \frac{n\upsilon_0^2\lambda^2}{2(1-\sqrt{2}\upsilon_0\lambda)}\right\}$$

Substituting this into (B.19), and following the proof of Bernstein's inequality (see, e.g., Theorem 2.10 in Boucheron, Lugosi and Massart [2013]), it can be shown that with probability at least $1 - \exp\{p\log(1+2/\varepsilon) - u\}$,

$$\sup_{\boldsymbol{\delta} \in \Theta(r)}\{D_h(\boldsymbol{\delta}) - \widehat{D}_h(\boldsymbol{\delta})\} \le \frac{2\sqrt{2}}{1-\varepsilon}\bar{\tau}\upsilon_0 r \cdot \left(\sqrt{\frac{u}{n}} + \frac{u}{n}\right). \tag{B.20}$$

This proves (B.15) by taking $\varepsilon = 2/(e^4 - 1)$.

Next, we prove the uniform bound (B.16), which holds for all $\boldsymbol{\delta} \in \Theta(r_l, r_u) := \{\boldsymbol{v} \in \mathbb{R}^p : r_l \le \|\boldsymbol{v}\|_{\boldsymbol{\Sigma}} \le r_u\}$, via a peeling argument. For some $\gamma > 1$ (to be specified) and positive integers $k = 1, \ldots, N := \lceil \log(\frac{r_u}{r_l})/\log(\gamma) \rceil$, define the sets $\Theta_k = \{\boldsymbol{v} \in \mathbb{R}^p : \gamma^{k-1} r_l \le \|\boldsymbol{v}\|_{\boldsymbol{\Sigma}} \le \gamma^k r_l\}$, so that

$\Theta(r_l, r_u) \subseteq \cup_{k=1}^{N} \Theta_k$. Then,

$$
\begin{aligned}
&\mathbb{P}\left\{ \exists \boldsymbol{\delta} \in \Theta(r_l, r_u) \text{ s.t. } D_h(\boldsymbol{\delta}) - \widehat{D}_h(\boldsymbol{\delta}) > \frac{2\sqrt{2}\gamma}{1-\varepsilon} \bar{\tau} \upsilon_0 \|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}} \cdot \left(\sqrt{\frac{u}{n}} + \frac{u}{n} \right) \right\} \\
&\leq \sum_{k=1}^{N} \mathbb{P}\left\{ \exists \boldsymbol{\delta} \in \Theta_k \text{ s.t. } D_h(\boldsymbol{\delta}) - \widehat{D}_h(\boldsymbol{\delta}) > \frac{2\sqrt{2}\gamma}{1-\varepsilon} \bar{\tau} \upsilon_0 \gamma^{k-1} r_l \cdot \left(\sqrt{\frac{u}{n}} + \frac{u}{n} \right) \right\} \\
&\leq \sum_{k=1}^{N} \mathbb{P}\left\{ \sup_{\boldsymbol{\delta} \in \Theta(\gamma^k r_l)} D_h(\boldsymbol{\delta}) - \widehat{D}_h(\boldsymbol{\delta}) > \frac{2\sqrt{2}}{1-\varepsilon} \bar{\tau} \upsilon_0 \gamma^k r_l \cdot \left(\sqrt{\frac{u}{n}} + \frac{u}{n} \right) \right\} \\
&\overset{\text{(i)}}{\leq} \sum_{k=1}^{N} \exp\left\{ p \log(1 + 2/\varepsilon) - u \right\} \leq \lceil \log(\tfrac{r_u}{r_l}) / \log(\gamma) \rceil \exp\left\{ p \log(1 + 2/\varepsilon) - u \right\},
\end{aligned}
$$

where inequality (i) is obtained by repeatedly using (B.20) with $r = \gamma^k r_l$ for $k = 1, \ldots, N$. Taking $\varepsilon = 2/(e^4 - 1)$ and $\gamma = e^{1/e}$ yields the claimed bound (B.16). □

B.2.3 An alternative proof to Theorem 3.3.1

From the proof of Theorem 3.3.1 in Section B.2.2, we see that the use of peeling argument will create an additional term $\log_2(1/h)$ in the upper bound, although it is further bounded by $\log(\log n)$ (under the prescribed constraint on h), a very slowly growing function of n. In this section, we argue that this extra term is an artifact of the proof technique, and can be avoided through a more careful analysis regarding the (local) restricted strong convexity of the empirical loss $\widehat{Q}_h(\cdot)$.

The key is to refine the convergence rate of $\widehat{\boldsymbol{\beta}}_h$ from $r_0 \asymp h$ to the claimed one, conditioned on $\widehat{\boldsymbol{\delta}} = \widehat{\boldsymbol{\beta}}_h - \boldsymbol{\beta}^* \in \Theta(r_0)$. The first-order optimality condition implies $\nabla \widehat{Q}_h(\widehat{\boldsymbol{\beta}}_h) = 0$, and hence

$$
\begin{aligned}
&\langle \nabla \widehat{Q}_h(\widehat{\boldsymbol{\beta}}_h) - \nabla \widehat{Q}_h(\boldsymbol{\beta}^*), \widehat{\boldsymbol{\delta}} \rangle \\
&= \langle -\nabla \widehat{Q}_h(\boldsymbol{\beta}^*), \widehat{\boldsymbol{\delta}} \rangle \leq \Big(\|\boldsymbol{\Sigma}^{-1/2} \{ \nabla \widehat{Q}_h(\boldsymbol{\beta}^*) - \nabla Q_h(\boldsymbol{\beta}^*) \}\|_2 + \underbrace{\|\boldsymbol{\Sigma}^{-1/2} \nabla Q_h(\boldsymbol{\beta}^*)\|_2}_{\leq 0.5 l_0 \kappa_2 \cdot h^2} \Big) \|\widehat{\boldsymbol{\delta}}\|_{\boldsymbol{\Sigma}}.
\end{aligned}
\tag{B.21}
$$

Define the symmetrized Bregman divergence associated with the convex function $\widehat{Q}_h(\cdot)$:

$$D(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = \langle \nabla \widehat{Q}_h(\boldsymbol{\beta}_1) - \nabla \widehat{Q}_h(\boldsymbol{\beta}_2), \boldsymbol{\beta}_1 - \boldsymbol{\beta}_2 \rangle \geq 0, \;\; \boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \mathbb{R}^p. \tag{B.22}$$

Then the left-hand side of (B.21) reads $D(\boldsymbol{\beta}^* + \widehat{\boldsymbol{\delta}}, \boldsymbol{\beta}^*)$. Starting from (B.21), we need to bound $\|\boldsymbol{\Sigma}^{-1/2}\{\nabla \widehat{Q}_h(\boldsymbol{\beta}^*) - \nabla Q_h(\boldsymbol{\beta}^*)\}\|_2$ from above, and derive a lower bound for $D(\boldsymbol{\beta}^* + \boldsymbol{\delta}, \boldsymbol{\beta}^*)$ uniformly over $\boldsymbol{\delta} \in \mathbb{R}^p$ in a local neighborhood of the origin. The following two lemmas serve for this purpose.

Lemma B.2.2. *Assume Conditions 3.3.1–3.3.4 hold. For any $t \geq 0$,*

$$\|\boldsymbol{\Sigma}^{-1/2}\{\nabla \widehat{Q}_h(\boldsymbol{\beta}^*) - \nabla Q_h(\boldsymbol{\beta}^*)\}\|_2 \leq 1.46 \upsilon_0 \left(C_\tau \sqrt{\frac{4p+2t}{n}} + \bar{\tau} \frac{4p+2t}{n} \right) \tag{B.23}$$

with probability at least $1 - e^{-t}$, where $C_\tau^2 = \tau(1-\tau) + (1+\tau) l_0 \kappa_2 h^2$.

In addition to Condition 3.3.2, assume $f_{\varepsilon|\boldsymbol{x}}(0) \leq \bar{f}$ for some $\bar{f} \geq \underline{f} > 0$. Then, for all $0 < h \leq \underline{f}/(4l_0)$,

$$\frac{7}{8}\underline{f} \leq \inf_{|u| \leq h/2} f_{\varepsilon|\boldsymbol{x}}(u) \leq \sup_{|u| \leq h/2} f_{\varepsilon|\boldsymbol{x}}(u) \leq \frac{9}{8}\bar{f}, \tag{B.24}$$

almost surely (over $\boldsymbol{x}$). Moreover, for $\delta \in (0, 1]$, define $\iota_\delta \geq 0$ as

$$\iota_\delta = \inf\big\{ \iota > 0 : \mathbb{E}\big\{ \langle \boldsymbol{u}, \boldsymbol{w} \rangle^2 \mathbb{1}\big(|\langle \boldsymbol{u}, \boldsymbol{w} \rangle| > \iota \big) \big\} \leq \delta \;\text{ for all } \boldsymbol{u} \in \mathbb{S}^{p-1} \big\}, \tag{B.25}$$

where $\boldsymbol{w} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}$ is the standardized covariate vector satisfying $\mathbb{E}(\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}}) = \mathbf{I}_p$, and hence $\mathbb{E}\langle \boldsymbol{u}, \boldsymbol{w} \rangle^2 = 1$ for any $\boldsymbol{u} \in \mathbb{S}^{p-1}$. It can be shown that ι_δ depends only on δ and υ_0 in Condition 3.3.4, and the map $\delta \mapsto \iota_\delta$ is non-increasing with $\iota_\delta \downarrow 0$ as $\delta \uparrow 1$ and $\iota_1 = 0$. By Markov's inequality, for any $\iota > 0$ it holds $\sup_{\boldsymbol{u} \in \mathbb{S}^{p-1}} \mathbb{E}\{ \langle \boldsymbol{u}, \boldsymbol{w} \rangle^2 \mathbb{1}(|\langle \boldsymbol{u}, \boldsymbol{w} \rangle| > \iota) \} \leq \iota^{-2} m_4$. Hence, a rather crude upper bound for ι_δ is $\iota_\delta \leq (m_4/\delta)^{1/2}$.

Lemma B.2.3. *Assume the kernel $K(\cdot)$ is such that $\kappa_l = \min_{|u|\le 1} K(u) > 0$, and let the bandwidth satisfy $0 < h \le \underline{f}/(4l_0)$. Given any $0 < r \le h/(4\iota_{1/4})$ with $\iota_{1/4}$ defined in* (B.25)*,*

$$\inf_{\boldsymbol{\delta}\in\Theta(r)} \frac{D(\boldsymbol{\beta}^* + \boldsymbol{\delta}, \boldsymbol{\beta}^*)}{\kappa_l \|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}}^2} \ge c_0 \underline{f} - c_1 \left\{ \bar{f}^{1/2} \frac{1}{r} \sqrt{\frac{ph}{n}} + (\bar{f} m_4)^{1/2} \sqrt{\frac{t}{nh}} + \frac{ht}{r^2 n} \right\} \tag{B.26}$$

with probability at least $1 - e^{-t}$ for any $t \ge 0$, where $c_1 > 0$ is an absolute constant, and $c_0 = 21/32$.

Let $\mathscr{G}(t)$ be the "good" event that the bounds (B.23) and (B.26) are satisfied. Together, Lemma B.2.2, Lemma B.2.3 and (B.21) imply that, conditioned on $\{\widehat{\boldsymbol{\delta}} \in \Theta(r_0)\} \cap \mathscr{G}(t)$ with $r_0 \le h/(4\iota_{1/4})$,

$$\begin{aligned} c_0 \kappa_l \underline{f} \cdot \|\widehat{\boldsymbol{\delta}}\|_{\boldsymbol{\Sigma}}^2 \le & \left\{ 1.46 \upsilon_0 \left(C_\tau \sqrt{\frac{4p+2t}{n}} + \bar{\tau} \frac{4p+2t}{n} \right) + 0.5 l_0 \kappa_2 h^2 \right\} \|\widehat{\boldsymbol{\delta}}\|_{\boldsymbol{\Sigma}} \\ & + c_1 \kappa_l \left\{ \bar{f}^{1/2} \frac{1}{r_0} \sqrt{\frac{ph}{n}} + (\bar{f} m_4)^{1/2} \sqrt{\frac{t}{nh}} + \frac{ht}{r_0^2 n} \right\} r_0 \cdot \|\widehat{\boldsymbol{\delta}}\|_{\boldsymbol{\Sigma}}. \end{aligned}$$

Consequently, we obtain the bound

$$\begin{aligned} c_0 \underline{f} \cdot \|\widehat{\boldsymbol{\delta}}\|_{\boldsymbol{\Sigma}} \le & 1.46 \kappa_l^{-1} \upsilon_0 \left(C_\tau \sqrt{\frac{4p+2t}{n}} + \bar{\tau} \frac{4p+2t}{n} \right) + 0.5 \kappa_l^{-1} l_0 \kappa_2 h^2 \\ & + c_1 \left\{ \bar{f}^{1/2} \sqrt{\frac{ph}{n}} + (\bar{f} m_4)^{1/2} r_0 \sqrt{\frac{t}{nh}} + \frac{ht}{r_0 n} \right\} \end{aligned}$$

without having the additional $\log_2(1/h)$ term. For example, we may take $r_0 = h/(8m_4^{1/2})$. □

Proof of Lemma B.2.2

Define $\xi_i = \mathscr{K}(-\varepsilon_i/h) - \tau$ for $i = 1, \ldots, n$, so that $\boldsymbol{\Sigma}^{-1/2}\{\nabla \widehat{Q}_h(\boldsymbol{\beta}^*) - \nabla Q_h(\boldsymbol{\beta}^*)\} = (1/n)\sum_{i=1}^n \{\xi_i \boldsymbol{w}_i - \mathbb{E}(\xi_i \boldsymbol{w}_i)\} \in \mathbb{R}^p$, where $\boldsymbol{w}_i = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}_i$. Using a covering argument, for any $\varepsilon \in (0,1)$, there exists an ε-net $\mathscr{N}_\varepsilon$ of the unit sphere with cardinality $|\mathscr{N}_\varepsilon| \le (1 + 2/\varepsilon)^p$ such

that

$$\|\boldsymbol{\Sigma}^{-1/2}\{\nabla\widehat{Q}_h(\boldsymbol{\beta}^*)-\nabla Q_h(\boldsymbol{\beta}^*)\}\|_2 \le (1-\varepsilon)^{-1}\max_{\boldsymbol{u}\in\mathscr{N}_\varepsilon}\big\langle \boldsymbol{u}, \boldsymbol{\Sigma}^{-1/2}\{\nabla\widehat{Q}_h(\boldsymbol{\beta}^*)-\nabla Q_h(\boldsymbol{\beta}^*)\}\big\rangle.$$

For each unit vector $\boldsymbol{u}\in\mathscr{N}_\varepsilon$, define centered random variables $\gamma_{\boldsymbol{u},i}=\langle \boldsymbol{u}, \xi_i\boldsymbol{w}_i-\mathbb{E}(\xi_i\boldsymbol{w}_i)\rangle$. It can be shown that $|\xi_i|\le\bar\tau:=\max(1-\tau,\tau)$ and $\mathbb{E}(\xi_i^2|\boldsymbol{x}_i)\le C_\tau=\tau(1-\tau)+(1+\tau)l_0\kappa_2h^2$. Hence, for $k=2,3,\ldots,$

$$\begin{aligned}
\mathbb{E}\big(|\langle \boldsymbol{u}, \xi_i\boldsymbol{w}_i\rangle|^k\big) &\le \bar\tau^{k-2}\,\mathbb{E}\big\{|\langle \boldsymbol{u}, \boldsymbol{w}_i\rangle|^k\cdot\mathbb{E}(\xi_i^2|\boldsymbol{x}_i)\big\}\\
&\le C_\tau^2\bar\tau^{k-2}\upsilon_0^k\int_0^\infty \mathbb{P}(|\langle \boldsymbol{u}, \boldsymbol{w}_i\rangle|\ge\upsilon_0 t)kt^{k-1}\,\mathrm{d}t\\
&\le C_\tau^2\bar\tau^{k-2}\upsilon_0^k k\int_0^\infty t^{k-1}e^{-t}\,\mathrm{d}t\\
&= k!\cdot C_\tau^2\bar\tau^{k-2}\cdot\upsilon_0^k\\
&\le \frac{k!}{2}\cdot(C_\tau\upsilon_0)^2\cdot(2\bar\tau\upsilon_0)^{k-2}.
\end{aligned}$$

Consequently, it follows from Bernstein's inequality that for every $u\ge0$,

$$\frac{1}{n}\sum_{i=1}^n\gamma_{\boldsymbol{u},i}\le\upsilon_0\left(C_\tau\sqrt{\frac{2u}{n}}+\frac{2\bar\tau u}{n}\right)$$

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

Finally, applying a union bound over $\boldsymbol{u}\in\mathscr{N}_\varepsilon$ yields

$$\|\boldsymbol{\Sigma}^{-1/2}\{\nabla\widehat{Q}_h(\boldsymbol{\beta}^*)-\nabla Q_h(\boldsymbol{\beta}^*)\}\|_2\le\frac{\upsilon_0}{1-\varepsilon}\left(C_\tau\sqrt{\frac{2u}{n}}+\frac{2\bar\tau u}{n}\right)$$

with probability at least $1-e^{\log(1+2/\varepsilon)p-u}$. Taking $\varepsilon=2/(e^2-1)$ and $u=2p+t$ ($t\ge0$) proves the claimed result. □

Proof of Lemma B.2.3

Recall that the empirical loss $\widehat{Q}_h(\cdot)$ in (3.5) is convex and twice continuously differentiable with

$$\nabla \widehat{Q}_h(\boldsymbol{\beta}) = (1/n)\sum_{i=1}^{n}\{\mathscr{K}_h(\langle \boldsymbol{x}_i, \boldsymbol{\beta}\rangle - y_i) - \tau\}\boldsymbol{x}_i \text{ and}$$
$$\nabla^2 \widehat{Q}_h(\boldsymbol{\beta}) = (1/n)\sum_{i=1}^{n} K_h(\langle \boldsymbol{x}_i, \boldsymbol{\beta}\rangle - y_i)\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}.$$

For the symmetrized Bregman divergence $D : \mathbb{R}^p \times \mathbb{R}^p \to [0,\infty)$ defined in (B.22), we have

$$D(\boldsymbol{\beta}^* + \boldsymbol{\delta}, \boldsymbol{\beta}^*) = \frac{1}{n}\sum_{i=1}^{n}\left\{\mathscr{K}\left(\frac{\langle \boldsymbol{w}_i, \boldsymbol{v}\rangle - \varepsilon_i}{h}\right) - \mathscr{K}\left(\frac{-\varepsilon_i}{h}\right)\right\}\langle \boldsymbol{w}_i, \boldsymbol{v}\rangle, \tag{B.27}$$

where $\boldsymbol{w}_i = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}_i$ and $\boldsymbol{v} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\delta}$. Define the events $\mathscr{E}_i = \{|\varepsilon_i| \le h/2\} \cap \{|\langle \boldsymbol{w}_i, \boldsymbol{v}\rangle| \le \|\boldsymbol{v}\|_2 \cdot h/(2r)\}$ for $i = 1, \ldots, n$. For any $\boldsymbol{v} \in \mathbb{B}^p(r)$, note that $|\varepsilon_i - \langle \boldsymbol{w}_i, \boldsymbol{v}\rangle| \le h$ on $\mathscr{E}_i$, implying

$$D(\boldsymbol{\beta}^* + \boldsymbol{\delta}, \boldsymbol{\beta}^*) \ge \frac{\kappa_l}{nh}\sum_{i=1}^{n}\langle \boldsymbol{w}_i, \boldsymbol{v}\rangle^2 \mathbb{1}_{\mathscr{E}_i}, \tag{B.28}$$

where $\mathbb{1}_{\mathscr{E}_i}$ is the indicator function of $\mathscr{E}_i$ and $\kappa_l = \min_{|u|\le 1} K(u)$. It then suffices to bound the right-hand side of the above inequality from below uniformly over $\boldsymbol{v} \in \mathbb{B}^p(r)$.

For $R > 0$, define the function $\varphi_R(u) = u^2\mathbb{1}(|u| \le R/2) + \{u\,\mathrm{sign}(u) - R\}^2\mathbb{1}(R/2 < |u| \le R)$, which is R-Lipschitz continuous and satisfies

$$u^2\mathbb{1}(|u| \le R/2) \le \varphi_R(u) \le u^2\mathbb{1}(|u| \le R). \tag{B.29}$$

Moreover, note that $\varphi_{cR}(cu) = c^2\varphi_R(u)$ for any $c > 0$ and $\varphi_0(u) = 0$. Hence,

$$\langle \boldsymbol{w}_i, \boldsymbol{v}\rangle^2 \mathbb{1}_{\mathscr{E}_i} \ge \varphi_{\|\boldsymbol{v}\|_2 h/(2r)}(\|\boldsymbol{v}\|_2\langle \boldsymbol{w}_i, \boldsymbol{v}/\|\boldsymbol{v}\|_2\rangle) \cdot \omega_i = \|\boldsymbol{v}\|_2^2 \cdot \varphi_{h/(2r)}(\langle \boldsymbol{w}_i, \boldsymbol{v}/\|\boldsymbol{v}\|_2\rangle) \cdot \omega_i,$$

where $\omega_i := \mathbb{1}(|\varepsilon_i| \le h/2)$. By a change of variable, the problem is reduced to bounding

$$D_0(\boldsymbol{v}) := \frac{1}{nh}\sum_{i=1}^{n} \omega_i \cdot \varphi_{h/(2r)}(\langle \boldsymbol{w}_i, \boldsymbol{v}\rangle) \tag{B.30}$$

from below uniformly over $\boldsymbol{v} \in \mathbb{S}^{p-1}$.

In the following, we bound the expectation $\mathbb{E}\{D_0(\boldsymbol{v})\}$ and the random fluctuation $D_0(\boldsymbol{v}) - \mathbb{E}\{D_0(\boldsymbol{v})\}$, separately, starting with the former. By (B.24),

$$\frac{7}{8}\underline{f}h \le \mathbb{E}(\omega_i|\boldsymbol{x}_i) = \int_{-h/2}^{h/2} f_{\varepsilon_i|\boldsymbol{x}_i}(u)\,\mathrm{d}u \le \frac{9}{8}\bar{f}h. \tag{B.31}$$

Moreover, define $\xi_{\boldsymbol{v}} = \langle \boldsymbol{w}, \boldsymbol{v}\rangle$ such that $\mathbb{E}(\xi_{\boldsymbol{v}}^2) = 1$. By (B.29) and (B.31),

$$\mathbb{E}\big\{\omega_i \cdot \varphi_{h/(2r)}(\langle \boldsymbol{w}_i, \boldsymbol{v}\rangle)\big\} \ge \frac{7}{8}\underline{f}h \cdot \mathbb{E}\varphi_{h/(2r)}(\langle \boldsymbol{w}_i, \boldsymbol{v}\rangle) \ge \frac{7}{8}\underline{f}h \cdot \big[1 - \mathbb{E}\xi_{\boldsymbol{v}}^2 \mathbb{1}\{|\xi_{\boldsymbol{v}}| > h/(4r)\}\big].$$

With $r \le h/(4\iota_{1/4})$ and $\iota_{1/4}$ defined in (B.25), it follows that

$$\inf_{\boldsymbol{v}\in\mathbb{S}^{p-1}} \mathbb{E}\{D_0(\boldsymbol{v})\} \ge \frac{7}{8}\underline{f} \cdot \left\{1 - \sup_{\boldsymbol{u}\in\mathbb{S}^{p-1}} \mathbb{E}\langle \boldsymbol{w}, \boldsymbol{u}\rangle^2 \mathbb{1}(|\langle \boldsymbol{w}, \boldsymbol{u}\rangle| \ge \iota_{1/4})\right\} \ge \frac{21}{32}\underline{f}. \tag{B.32}$$

Turning to the random fluctuation, we will use Theorem 7.3 in Bousquet [2003] (a refined Talagrand's inequality) to bound

$$\Delta = \sup_{\boldsymbol{v}\in\mathbb{S}^{p-1}} \big\{D_0^-(\boldsymbol{v}) - \mathbb{E}D_0^-(\boldsymbol{v})\big\}, \tag{B.33}$$

where $D_0^-(\boldsymbol{v}) := -D_0(\boldsymbol{v})$. Note that $0 \le \varphi_R(u) \le (R/2)^2$ for all $u \in \mathbb{R}$ and $\omega_i \in \{0,1\}$. Hence, $\chi_i := (\omega_i/h) \cdot \varphi_{h/(2r)}(\langle \boldsymbol{w}_i, \boldsymbol{v}\rangle) \ge 0$ is bounded by $h/(4r)^2$. Moreover, it follows from (B.31) that $\mathbb{E}(\chi_i^2) \le 9\bar{f}m_4/(8h)$. We then apply Theorem 7.3 in Bousquet [2003] and obtain that, for any

$t > 0$,

$$\Delta \le \mathbb{E}(\Delta) + \{\mathbb{E}(\Delta)\}^{1/2} \frac{1}{2r}\sqrt{\frac{ht}{n}} + \frac{3}{2}(\bar{f}m_4)^{1/2}\sqrt{\frac{t}{nh}} + \frac{h}{(4r)^2}\frac{t}{3n} \tag{B.34}$$

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

It remains to bound the expected value $\mathbb{E}(\Delta)$. Recall that $\omega_i = \mathbb{1}(|\varepsilon_i| \le h/2) \in \{0,1\}$, and hence $\omega_i \varphi_{h/(2r)}(\langle \boldsymbol{w}_i, \boldsymbol{v}\rangle) = \omega_i^2 \varphi_{h/(2r)}(\langle \boldsymbol{w}_i, \boldsymbol{v}\rangle) = \varphi_{\omega_i h/(2r)}(\langle \omega_i \boldsymbol{w}_i, \boldsymbol{v}\rangle) = \varphi_{h/(2r)}(\langle \omega_i \boldsymbol{w}_i, \boldsymbol{v}\rangle)$. Define

$$\bar{\boldsymbol{w}}_i = \omega_i \boldsymbol{w}_i \quad \text{and} \quad \mathscr{E}(\boldsymbol{v}; \bar{\boldsymbol{w}}_i) = \varphi_{h/(2r)}(\langle \bar{\boldsymbol{w}}_i, \boldsymbol{v}\rangle), \ \ \boldsymbol{v} \in \mathbb{S}^{p-1}.$$

By Rademacher symmetrization,

$$\mathbb{E}(\Delta) \le 2\mathbb{E}\left\{ \sup_{\boldsymbol{v}\in\mathbb{S}^{p-1}} \frac{1}{nh}\sum_{i=1}^{n} e_i \cdot \mathscr{E}(\boldsymbol{v}; \bar{\boldsymbol{w}}_i) \right\},$$

where $e_1, \ldots, e_n$ are independent Rademacher random variables. Since $\varphi_R(\cdot)$ is R-Lipschitz, $\mathscr{E}(\boldsymbol{v}; \boldsymbol{z}_i)$ is an $(h/2r)$-Lipschitz function in $\langle \bar{\boldsymbol{w}}_i, \boldsymbol{v}\rangle$, i.e., for any $\bar{\boldsymbol{w}}_i$ and parameters $\boldsymbol{v}, \boldsymbol{v}' \in \mathbb{S}^{p-1}$,

$$\left|\mathscr{E}(\boldsymbol{v}; \bar{\boldsymbol{w}}_i) - \mathscr{E}(\boldsymbol{v}'; \bar{\boldsymbol{w}}_i)\right| \le \frac{h}{2r}\left|\langle \bar{\boldsymbol{w}}_i, \boldsymbol{v}\rangle - \langle \bar{\boldsymbol{w}}_i, \boldsymbol{v}'\rangle\right|. \tag{B.35}$$

Moreover, observe that $\mathscr{E}(\boldsymbol{v}; \bar{\boldsymbol{w}}_i) = 0$ for any $\boldsymbol{v}$ such that $\langle \bar{\boldsymbol{w}}_i, \boldsymbol{v}\rangle = 0$. With the above preparations, we are ready to use Talagrand's contraction principle to bound $\mathbb{E}(\Delta)$. Define the subset $T \subseteq \mathbb{R}^n$ as

$$T = \left\{ \boldsymbol{t} = (t_1, \ldots, t_n)^{\mathrm{T}} : t_i = \langle \bar{\boldsymbol{w}}_i, \boldsymbol{v}\rangle, i = 1, \ldots, n, \boldsymbol{v} \in \mathbb{S}^{p-1} \right\},$$

and contractions $\phi_i : \mathbb{R} \to \mathbb{R}$ as $\phi_i(t) = (2r/h) \cdot \varphi_{h/(2r)}(t)$. By (B.35), $|\phi(t) - \phi(s)| \le |t - s|$ for all $t, s \in \mathbb{R}$. Applying Talagrand's contraction principle (see, e.g., Theorem 4.12 and (4.20) in

Ledoux and Talagrand [1991]), we have

$$\mathbb{E}(\Delta) \leq 2\mathbb{E}\left\{\sup_{\boldsymbol{v}\in\mathbb{S}^{p-1}} \frac{1}{nh}\sum_{i=1}^{n} e_i \cdot \mathscr{E}(\boldsymbol{v};\bar{\boldsymbol{w}}_i)\right\} = \mathbb{E}\left\{\sup_{\boldsymbol{t}\in T} \frac{1}{nr}\sum_{i=1}^{n} e_i\phi_i(t_i)\right\} \leq \mathbb{E}\left(\sup_{\boldsymbol{t}\in T} \frac{1}{nr}\sum_{i=1}^{n} e_i t_i\right)$$
$$= \mathbb{E}\left(\sup_{\boldsymbol{v}\in\mathbb{S}^{p-1}} \frac{1}{nr}\sum_{i=1}^{n} e_i\langle\bar{\boldsymbol{w}}_i,\boldsymbol{v}\rangle\right) \leq \mathbb{E}\left\|\frac{1}{nr}\sum_{i=1}^{n} e_i\bar{\boldsymbol{w}}_i\right\|_2 \leq \bar{f}^{1/2}\frac{3}{2r}\sqrt{\frac{ph}{2n}}.$$

This, combined with (B.33) and (B.34), yields that

$$\Delta \leq \frac{3}{2}\bar{f}^{1/2}\left(1.25\frac{1}{r}\sqrt{\frac{ph}{2n}} + m_4^{1/2}\sqrt{\frac{t}{nh}}\right) + (1+1/48)\frac{ht}{r^2 n}$$

with probability at least $1-e^{-t}$. Combining this with (B.27), (B.30) and (B.32) proves (B.26).

□

B.2.4 Proof of Theorem 3.3.2

We keep the notation used in the proof of Theorem 3.3.1, and for any $t \geq 0$, let $r = r(n,p,t) \asymp \sqrt{(p+t)/n} + h^2 > 0$ be such that $\mathbb{P}\{\widehat{\boldsymbol{\beta}}_h \in \boldsymbol{\beta}^* + \Theta(r)\} \geq 1-2e^{-t}$, provided $\sqrt{(p+t)/n} \lesssim h \lesssim 1$. Define the vector-valued random process

$$\Delta(\boldsymbol{\delta}) = \boldsymbol{\Sigma}^{-1/2}\big\{\nabla\widehat{Q}_h(\boldsymbol{\beta}^*+\boldsymbol{\delta}) - \nabla\widehat{Q}_h(\boldsymbol{\beta}^*) - \mathbf{J}_h\boldsymbol{\delta}\big\},\ \ \boldsymbol{\delta}\in\mathbb{R}^p, \tag{B.36}$$

where $\mathbf{J}_h = \nabla^2 Q_h(\boldsymbol{\beta}^*)$ is the population Hessian at $\boldsymbol{\beta}^*$. Since $\widehat{\boldsymbol{\beta}}_h$ falls in a local neighborhood of $\boldsymbol{\beta}^*$ with high probability, it suffices to bound the local fluctuation $\sup_{\boldsymbol{\delta}\in\Theta(r)}\|\Delta(\boldsymbol{\delta})\|_2$. By the triangle inequality,

$$\sup_{\boldsymbol{\delta}\in\Theta(r)}\|\Delta(\boldsymbol{\delta})\|_2 \leq \sup_{\boldsymbol{\delta}\in\Theta(r)}\|\mathbb{E}\Delta(\boldsymbol{\delta})\|_2 + \sup_{\boldsymbol{\delta}\in\Theta(r)}\|\Delta(\boldsymbol{\delta})-\mathbb{E}\Delta(\boldsymbol{\delta})\|_2 := I_1 + I_2. \tag{B.37}$$

We now provide upper bounds for I_1 and I_2, respectively.

UPPER BOUND FOR I_1: By the mean value theorem for vector-valued functions,

$$\begin{aligned}\mathbb{E}\Delta(\boldsymbol{\delta}) &= \boldsymbol{\Sigma}^{-1/2}\left\langle \int_0^1 \nabla^2 Q_h\big(\boldsymbol{\beta}^* + t\boldsymbol{\delta}\big)\mathrm{d}t, \boldsymbol{\delta}\right\rangle - \boldsymbol{\Sigma}^{-1/2}\mathbf{J}_h\boldsymbol{\delta} \\ &= \left\langle \boldsymbol{\Sigma}^{-1/2}\int_0^1 \nabla^2 Q_h\big(\boldsymbol{\beta}^* + t\boldsymbol{\delta}\big)\mathrm{d}t\,\boldsymbol{\Sigma}^{-1/2} - \mathbf{H}_h, \boldsymbol{\Sigma}^{1/2}\boldsymbol{\delta}\right\rangle,\end{aligned}$$

where $\mathbf{H}_h := \boldsymbol{\Sigma}^{-1/2}\mathbf{J}_h\boldsymbol{\Sigma}^{-1/2} = \mathbb{E}\{K_h(\varepsilon)\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}}\}$. By law of iterative expectation and a change of variable,

$$\begin{aligned}\boldsymbol{\Sigma}^{-1/2}\nabla^2 Q_h(\boldsymbol{\beta}^* + t\boldsymbol{\delta})\boldsymbol{\Sigma}^{-1/2} &= \mathbb{E}\big\{K_h(t\langle \boldsymbol{x}, \boldsymbol{\delta}\rangle - \varepsilon)\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}}\big\} \\ &= \mathbb{E}\left\{\int_{-\infty}^{\infty} K(u) f_{\varepsilon|\boldsymbol{x}}(t\langle \boldsymbol{x}, \boldsymbol{\delta}\rangle - hu)\,\mathrm{d}u \cdot \boldsymbol{w}\boldsymbol{w}^{\mathrm{T}}\right\}.\end{aligned}$$

Write $\boldsymbol{v} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\delta}$ for $\boldsymbol{\delta} \in \Theta(r)$, so that $\|\boldsymbol{v}\|_2 \le r$ and

$$\boldsymbol{\Sigma}^{-1/2}\nabla^2 Q_h\big(\boldsymbol{\beta}^* + t\boldsymbol{\delta}\big)\boldsymbol{\Sigma}^{-1/2} = \mathbb{E}\left\{\int_{-\infty}^{\infty} K(u) f_{\varepsilon|\boldsymbol{x}}(t\langle \boldsymbol{w}, \boldsymbol{v}\rangle - hu)\,\mathrm{d}u \cdot \boldsymbol{w}\boldsymbol{w}^{\mathrm{T}}\right\}.$$

By the Lipschitz continuity of $f_{\varepsilon|\boldsymbol{x}}(\cdot)$,

$$\begin{aligned}&\big\|\boldsymbol{\Sigma}^{-1/2}\nabla^2 Q_h\big(\boldsymbol{\beta}^* + t\boldsymbol{\delta}\big)\boldsymbol{\Sigma}^{-1/2} - \mathbf{H}_h\big\|_2 \\ &= \left\|\mathbb{E}\int K(u)\big\{f_{\varepsilon|\boldsymbol{x}}(t\langle \boldsymbol{w}, \boldsymbol{v}\rangle - hu) - f_{\varepsilon|\boldsymbol{x}}(-hu)\big\}\,\mathrm{d}u \cdot \boldsymbol{w}\boldsymbol{w}^{\mathrm{T}}\right\|_2 \\ &\le l_0 t \sup_{\|\boldsymbol{u}\|_2=1} \mathbb{E}\big(\langle \boldsymbol{w}, \boldsymbol{u}\rangle^2 |\langle \boldsymbol{w}, \boldsymbol{v}\rangle|\big) \le l_0 m_3 r t,\end{aligned}$$

where the third inequality holds by the Cauchy-Schwarz inequality. Consequently,

$$\sup_{\boldsymbol{\delta}\in\Theta(r)} \|\mathbb{E}\Delta(\boldsymbol{\delta})\|_2 \le 0.5 l_0 m_3 r^2. \tag{B.38}$$

UPPER BOUND FOR I_2: Next, we provide an upper bound for the supremum of the zero-

mean stochastic process $\Delta(\boldsymbol{\delta}) - \mathbb{E}\Delta(\boldsymbol{\delta})$ under ℓ_2-norm. Define the centered gradient process $G(\boldsymbol{\beta}) = \boldsymbol{\Sigma}^{-1/2}\{\nabla \widehat{Q}_h(\boldsymbol{\beta}) - \nabla Q_h(\boldsymbol{\beta})\}$, so that $\Delta(\boldsymbol{\delta}) - \mathbb{E}\Delta(\boldsymbol{\delta}) = G(\boldsymbol{\beta}^* + \boldsymbol{\delta}) - G(\boldsymbol{\beta}^*)$. Again, by a change of variable $\boldsymbol{v} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\delta}$,

$$\begin{aligned}\sup_{\boldsymbol{\delta}\in\Theta(r)} \|\Delta(\boldsymbol{\beta}) - \mathbb{E}\Delta(\boldsymbol{\beta})\|_2 &\le \sup_{\boldsymbol{\delta}\in\Theta(r)} \|G(\boldsymbol{\beta}^* + \boldsymbol{\delta}) - G(\boldsymbol{\beta}^*)\|_2 \\ &= \sup_{\|\boldsymbol{v}\|_2\le r} \|\underbrace{G(\boldsymbol{\beta}^* + \boldsymbol{\Sigma}^{-1/2}\boldsymbol{v}) - G(\boldsymbol{\beta}^*)}_{=:\Delta_0(\boldsymbol{v})}\|_2.\end{aligned}$$

We will employ Theorem A.3 in Spokoiny [2013] to bound the supremum $\sup_{\|\boldsymbol{v}\|_2\le r} \|\Delta_0(\boldsymbol{v})\|_2$, where $\Delta_0(\cdot)$ defined above satisfies $\Delta_0(\mathbf{0}) = \mathbf{0}$ and $\mathbb{E}\{\Delta_0(\boldsymbol{v})\} = \mathbf{0}$. Taking the gradient with respect to $\boldsymbol{v}$ yields

$$\nabla\Delta_0(\boldsymbol{v}) = \frac{1}{n}\sum_{i=1}^n \big\{K_{i,\boldsymbol{v}}\boldsymbol{w}_i\boldsymbol{w}_i^{\mathrm{T}} - \mathbb{E}\big(K_{i,\boldsymbol{v}}\boldsymbol{w}_i\boldsymbol{w}_i^{\mathrm{T}}\big)\big\},$$

where $K_{i,\boldsymbol{v}} := K_h(\langle \boldsymbol{w}_i, \boldsymbol{v}\rangle - \varepsilon_i)$ satisfies $0 \le K_{i,\boldsymbol{v}} \le \kappa_u h^{-1}$. For any $\boldsymbol{u}, \boldsymbol{u}' \in \mathbb{S}^{p-1}$ and $\lambda \in \mathbb{R}$, using the elementary inequality $|e^u - 1 - u| \le u^2 e^{|u|}/2$, we obtain

$$\begin{aligned}&\mathbb{E}\exp\big\{\lambda n^{1/2}\langle \boldsymbol{u}, \nabla\Delta_0(\boldsymbol{v})\boldsymbol{u}'\rangle/\upsilon_1^2\big\} \\ &\le \prod_{i=1}^n \Bigg\{1 + \frac{\lambda^2}{2\upsilon_1^4 n} e^{\frac{\bar{f}|\lambda|}{\upsilon_1^2\sqrt{n}}\mathbb{E}|\langle \boldsymbol{w}_i,\boldsymbol{u}\rangle\langle \boldsymbol{w}_i,\boldsymbol{u}'\rangle|} \times \\ &\qquad\qquad \mathbb{E}\big\{K_{i,\boldsymbol{v}}\langle \boldsymbol{w}_i,\boldsymbol{u}\rangle\langle \boldsymbol{w}_i,\boldsymbol{u}'\rangle - \mathbb{E}(K_{i,\boldsymbol{v}}\langle \boldsymbol{w}_i,\boldsymbol{u}\rangle\langle \boldsymbol{w}_i,\boldsymbol{u}'\rangle)\big\}^2 e^{\frac{\kappa_u|\lambda|}{h\sqrt{n}}|\langle \boldsymbol{w}_i,\boldsymbol{u}\rangle\langle \boldsymbol{w}_i,\boldsymbol{u}'\rangle|/\upsilon_1^2}\Bigg\} \\ &\le \prod_{i=1}^n \Bigg\{1 + \frac{\lambda^2}{2\upsilon_1^4 n} e^{\frac{\bar{f}|\lambda|}{\sqrt{n}}} \mathbb{E}\big\{K_{i,\boldsymbol{v}}\langle \boldsymbol{w}_i,\boldsymbol{u}\rangle\langle \boldsymbol{w}_i,\boldsymbol{u}'\rangle - \mathbb{E}(K_{i,\boldsymbol{v}}\langle \boldsymbol{w}_i,\boldsymbol{u}\rangle\langle \boldsymbol{w}_i,\boldsymbol{u}'\rangle)\big\}^2 e^{\frac{\kappa_u|\lambda|}{h\sqrt{n}}|\langle \boldsymbol{w}_i,\boldsymbol{u}\rangle\langle \boldsymbol{w}_i,\boldsymbol{u}'\rangle|/\upsilon_1^2}\Bigg\},\end{aligned} \tag{B.39}$$

where we used the bound $\mathbb{E}|\langle \boldsymbol{w}_i,\boldsymbol{u}\rangle\langle \boldsymbol{w}_i,\boldsymbol{u}'\rangle| \le (\mathbb{E}\langle \boldsymbol{w}_i,\boldsymbol{u}\rangle^2)^{1/2}(\mathbb{E}\langle \boldsymbol{w}_i,\boldsymbol{u}'\rangle^2)^{1/2} = 1$ in the second inequality. Moreover, the first and second conditional moments of $K_{i,\boldsymbol{v}}$ can be rewritten as

follows:

$$\mathbb{E}\big(K_{i,\boldsymbol{v}}|\boldsymbol{x}_i\big) = \frac{1}{h}\int_{-\infty}^{\infty} K\left(\frac{\langle \boldsymbol{w}_i,\boldsymbol{v}\rangle - t}{h}\right) f_{\varepsilon_i|\boldsymbol{x}_i}(t)\,\mathrm{d}t = \int_{-\infty}^{\infty} K(u) f_{\varepsilon_i|\boldsymbol{x}_i}(\langle \boldsymbol{w}_i,\boldsymbol{v}\rangle - hu)\,\mathrm{d}u;$$
$$\mathbb{E}\big(K^2_{i,\boldsymbol{v}}|\boldsymbol{x}_i\big) = \frac{1}{h^2}\int_{-\infty}^{\infty} K^2\left(\frac{\langle \boldsymbol{w}_i,\boldsymbol{v}\rangle - t}{h}\right) f_{\varepsilon_i|\boldsymbol{x}_i}(t)\,\mathrm{d}t = \frac{1}{h}\int_{-\infty}^{\infty} K^2(u) f_{\varepsilon_i|\boldsymbol{x}_i}(\langle \boldsymbol{w}_i,\boldsymbol{v}\rangle - hu)\,\mathrm{d}u,$$

from which it follows that $|\mathbb{E}(K_{i,\boldsymbol{v}}|\boldsymbol{x}_i)| \le \bar{f}$ and $\mathbb{E}(K^2_{i,\boldsymbol{v}}|\boldsymbol{x}_i) \le \kappa_u \bar{f} h^{-1}$ almost surely.

By the Cauchy-Schwarz inequality and the inequality $ab \le a^2/2 + b^2/2$, $a,b \in \mathbb{R}$, we have

$$\begin{aligned}
&\mathbb{E}(\langle \boldsymbol{w}_i,\boldsymbol{u}\rangle\langle \boldsymbol{w}_i,\boldsymbol{u}'\rangle)^2 e^{t|\langle \boldsymbol{w}_i,\boldsymbol{u}\rangle\langle \boldsymbol{w}_i,\boldsymbol{u}'\rangle|} \\
&\le \mathbb{E}(\langle \boldsymbol{w}_i,\boldsymbol{u}\rangle\langle \boldsymbol{w}_i,\boldsymbol{u}'\rangle)^2 e^{\frac{t}{2}\langle \boldsymbol{w}_i,\boldsymbol{u}\rangle^2 + \frac{t}{2}\langle \boldsymbol{w}_i,\boldsymbol{u}'\rangle^2} \\
&\le \Big(\mathbb{E}\langle \boldsymbol{w}_i,\boldsymbol{u}\rangle^4 e^{t\langle \boldsymbol{w}_i,\boldsymbol{u}\rangle^2}\Big)^{1/2}\Big(\mathbb{E}\langle \boldsymbol{w}_i,\boldsymbol{u}'\rangle^4 e^{t\langle \boldsymbol{w}_i,\boldsymbol{u}'\rangle^2}\Big)^{1/2}, \ \text{ valid for any } t>0.
\end{aligned}$$

Given a unit vector $\boldsymbol{u}$, let $\chi = \langle \boldsymbol{w},\boldsymbol{u}\rangle^2/(2\upsilon_1)^2$ so that under Condition 3.3.5, $\mathbb{P}(\chi \ge u) \le 2e^{-2u}$ for any $u \ge 0$. It follows that $\mathbb{E}(e^{\chi}) = 1 + \int_0^{\infty} e^u \mathbb{P}(\chi \ge u)\mathrm{d}u \le 1 + 2\int_0^{\infty} e^{-u}\mathrm{d}u = 3$, and

$$\mathbb{E}(\chi^2 e^{\chi}) = \int_0^{\infty} (u^2+2u)e^u \mathbb{P}(\chi \ge u)\mathrm{d}u \le 2\int_0^{\infty}(u^2+2u)e^{-u}\mathrm{d}u = 8.$$

Taking the supremum over $\boldsymbol{u} \in \mathbb{S}^{p-1}$, we have

$$\sup_{\boldsymbol{u}\in\mathbb{S}^{p-1}} \mathbb{E}e^{\langle \boldsymbol{w},\boldsymbol{u}\rangle^2/(2\upsilon_1)^2} \le 3 \ \text{ and } \ \sup_{\boldsymbol{u}\in\mathbb{S}^{p-1}} \mathbb{E}\langle \boldsymbol{w},\boldsymbol{u}\rangle^4 e^{\langle \boldsymbol{w},\boldsymbol{u}\rangle^2/(2\upsilon_1)^2} \le 8(2\upsilon_1)^4.$$

Substituting the above bounds into (B.39) yields that, for any $|\lambda| \le \min\{h/(4\kappa_u), 1/\bar{f}\}n^{1/2}$,

$$
\begin{aligned}
&\mathbb{E}\exp\big\{\lambda n^{1/2}\langle \boldsymbol{u}, \nabla\Delta_0(\boldsymbol{v})\boldsymbol{u}'\rangle/\upsilon_1^2\big\} \\
&\le \prod_{i=1}^n \Bigg[1 + \frac{e\lambda^2}{2\upsilon_1^4 n}\mathbb{E}\big\{K_{i,\boldsymbol{v}}\langle \boldsymbol{w}_i,\boldsymbol{u}\rangle\langle \boldsymbol{w}_i,\boldsymbol{u}'\rangle - \mathbb{E}(K_{i,\boldsymbol{v}}\langle \boldsymbol{w}_i,\boldsymbol{u}\rangle\langle \boldsymbol{w}_i,\boldsymbol{u}'\rangle)\big\}^2 e^{|\langle \boldsymbol{w}_i,\boldsymbol{u}\rangle\langle \boldsymbol{w}_i,\boldsymbol{u}'\rangle|/(4\upsilon_1^2)}\Bigg] \\
&\le \prod_{i=1}^n \Bigg[1 + \frac{e\lambda^2}{\upsilon_1^4 n}\mathbb{E}\big(K_{i,\boldsymbol{v}}\langle \boldsymbol{w}_i,\boldsymbol{u}\rangle\langle \boldsymbol{w}_i,\boldsymbol{u}'\rangle\big)^2 e^{|\langle \boldsymbol{w}_i,\boldsymbol{u}\rangle\langle \boldsymbol{w}_i,\boldsymbol{u}'\rangle|/(4\upsilon_1^2)} \\
&\qquad\qquad + \frac{e\lambda^2}{\upsilon_1^4 n}\big\{\mathbb{E}\big(K_{i,\boldsymbol{v}}\langle \boldsymbol{w}_i,\boldsymbol{u}\rangle\langle \boldsymbol{w}_i,\boldsymbol{u}'\rangle\big)\big\}^2 \mathbb{E}e^{|\langle \boldsymbol{w}_i,\boldsymbol{u}\rangle\langle \boldsymbol{w}_i,\boldsymbol{u}'\rangle|/(4\upsilon_1^2)}\Bigg] \\
&\le \prod_{i=1}^n \left(1 + C_0^2\frac{\lambda^2}{2nh}\right) \le \exp\big\{C_0^2\lambda^2/(2h)\big\},
\end{aligned}
$$

where $C_0 > 0$ depends only on $(\kappa_u, \bar{f})$. We have thus verified condition (A.4) in Spokoiny [2013] with $\mathrm{g} = \min\{h/(4\kappa_u), 1/\bar{f}\}(n/2)^{1/2}$ and $\nu_0 = C_0 h^{-1/2}$. Applying Theorem A.3 therein to the process $\{\Delta_0(\boldsymbol{r})/\upsilon_1^2, \boldsymbol{v} \in \mathbb{B}^p(r)\}$, we obtain that with probability at least $1 - e^{-t}$,

$$
\sup_{\|\boldsymbol{v}\|_2 \le r} \|\Delta_0(\boldsymbol{v})\|_2 \le 6C_0\upsilon_1^2 r\sqrt{\frac{4p+2t}{nh}}
$$

as long as $h \ge 8\kappa_u\sqrt{(2p+t)/n}$ and $n \ge 4\bar{f}^2(2p+t)$.

Joint with (B.37) and (B.38), this implies that with probability at least $1 - e^{-t}$,

$$
\sup_{\boldsymbol{\delta}\in\Theta(r)} \|\Delta(\boldsymbol{\delta})\|_2 \le 6C_0\upsilon_1^2 r\sqrt{\frac{4p+2t}{nh}} + 0.5l_0 m_3 r^2. \tag{B.40}
$$

Recall from the beginning of the proof that $\widehat{\boldsymbol{\delta}} = \widehat{\boldsymbol{\beta}}_h - \boldsymbol{\beta}^* \in \Theta(r)$ with probability at least $1 - 2e^{-t}$ with $r = r(n,p,t) \asymp \sqrt{(p+t)/n} + h^2$. Combined with (B.40), we conclude that with probability at least $1 - 3e^{-t}$, $\|\Delta(\widehat{\boldsymbol{\delta}})\|_2 \lesssim (p+t)/(h^{1/2}n) + h^{3/2}\sqrt{(p+t)/n} + h^4$, as claimed. □

B.2.5 Proof of Theorem 3.3.3

Let $\boldsymbol{a} \in \mathbb{R}^p$ be an arbitrary vector defining a linear functional of interest. Given $h = h_n > 0$, define $S_n = n^{-1/2}\sum_{i=1}^n \gamma_i \xi_i$ and its centered version $S_n^0 = S_n - \mathbb{E}(S_n)$, where $\xi_i = \tau - \mathscr{K}(-\varepsilon_i/h)$ and $\gamma_i = \langle \mathbf{J}_h^{-1}\boldsymbol{a}, \boldsymbol{x}_i\rangle$. By the Lipschitz continuity of $f_{\varepsilon|\boldsymbol{x}}(\cdot)$ and the fundamental theorem of calculus, it can be shown that $|\mathbb{E}(\xi_i|\boldsymbol{x}_i)| \le 0.5 l_0 \kappa_2 h^2$, from which it follows by the law of iterated expectation that $|\mathbb{E}(\gamma_i\xi_i)| \le 0.5 l_0 \kappa_2 \|\mathbf{J}_h^{-1}\boldsymbol{a}\|_{\boldsymbol{\Sigma}} \cdot h^2$.

Let $\eta_n = (p + \log n)/n$. Then, applying (B.36) and (B.40) with $t = \log n$ and the triangle inequality, we obtain that under the constraint $\eta_n^{1/2} \lesssim h \lesssim 1$,

$$
\begin{aligned}
&\big|n^{1/2}\langle \boldsymbol{a}, \widehat{\boldsymbol{\beta}}_h - \boldsymbol{\beta}^*\rangle - S_n^0\big| \\
&= n^{1/2}\left|\left\langle \boldsymbol{\Sigma}^{1/2}\mathbf{J}_h^{-1}\boldsymbol{a}, \boldsymbol{\Sigma}^{-1/2}\mathbf{J}_h(\widehat{\boldsymbol{\beta}}_h - \boldsymbol{\beta}^*) - \boldsymbol{\Sigma}^{-1/2}\frac{1}{n}\sum_{i=1}^n \big\{\tau - \mathscr{K}(-\varepsilon_i/h)\big\}\boldsymbol{x}_i \right\rangle\right| + |\mathbb{E}(S_n)| \\
&\le c_1 \|\mathbf{J}_h^{-1}\boldsymbol{a}\|_{\boldsymbol{\Sigma}} \cdot n^{1/2}\big(h^{-1/2}\delta_n + h^2\big) \qquad \text{(B.41)}
\end{aligned}
$$

with probability at least $1 - 3n^{-1}$ for some constant $c_1 > 0$.

For the centered partial sum $S_n^0 = S_n - \mathbb{E}(S_n) = n^{-1/2}\sum_{i=1}^n (1-\mathbb{E})\gamma_i\xi_i$, we have $\mathrm{var}(S_n^0) = \mathrm{var}(S_n) = \mathbb{E}(\gamma\xi)^2 - \{\mathbb{E}(\gamma\xi)\}^2$, where $\gamma = \langle \mathbf{J}_h^{-1}\boldsymbol{a}, \boldsymbol{x}\rangle$ and $\xi = \tau - \mathscr{K}(-\varepsilon/h)$. By the Berry-Esseen inequality (see, e.g., Tyurin [2011]),

$$
\sup_{x\in\mathbb{R}}\big|\mathbb{P}\big\{S_n^0 \le \mathrm{var}(S_n)^{1/2}x\big\} - \Phi(x)\big| \le \frac{\mathbb{E}\{|\gamma\xi - \mathbb{E}(\gamma\xi)|^3\}}{2[\mathbb{E}(\gamma\xi)^2 - \{\mathbb{E}(\gamma\xi)\}^2]^{3/2}\sqrt{n}}. \qquad \text{(B.42)}
$$

We have shown that $|\mathbb{E}(\gamma\xi)| \lesssim \|\mathbf{J}_h^{-1}\boldsymbol{a}\|_{\boldsymbol{\Sigma}} \cdot h^2$. To bound $\mathbb{E}(\xi^2|\boldsymbol{x})$, by a change of variable and

integration by parts,

$$\begin{aligned}\mathbb{E}\{\mathscr{K}^2(-\varepsilon/h)|\boldsymbol{x}\} &= 2\int_{-\infty}^{\infty} K(v)\mathscr{K}(v)F_{\varepsilon|\boldsymbol{x}}(-vh)\mathrm{d}v \\ &= 2\tau\underbrace{\int_{-\infty}^{\infty} K(v)\mathscr{K}(v)\mathrm{d}v}_{=1/2} - 2hf_{\varepsilon|\boldsymbol{x}}(0)\underbrace{\int_{-\infty}^{\infty} vK(v)\mathscr{K}(v)\mathrm{d}v}_{=\int_0^\infty \mathscr{K}(v)\{1-\mathscr{K}(v)\}\mathrm{d}v>0} \\ &\quad + 2\int_{-\infty}^{\infty}\int_{0}^{-vh}\{f_{\varepsilon|\boldsymbol{x}}(t)-f_{\varepsilon|\boldsymbol{x}}(0)\}K(v)\mathscr{K}(v)\mathrm{d}t\mathrm{d}v \\ &\leq \tau + l_0\kappa_2 h^2,\end{aligned}$$

where κ_2 and l_0 are the constants from Conditions 3.3.1 and 3.3.2. It then follows that

$$\tau(1-\tau) - Ch - (1+\tau)l_0\kappa_2 h^2 \leq \mathbb{E}(\xi^2|\boldsymbol{x}_i) \leq \tau(1-\tau) + (1+\tau)l_0\kappa_2 h^2,$$

where $C = 2\bar{f}\cdot\int_0^\infty \mathscr{K}(v)\{1-\mathscr{K}(v)\}\mathrm{d}v$. For all sufficiently small h, $\mathrm{var}(S_n) = \{\tau(1-\tau) + O(h)\}\|\mathbf{J}_h^{-1}\boldsymbol{a}\|_{\boldsymbol{\Sigma}}^2$. On the other hand,

$$\mathbb{E}(|\gamma\xi|^3) \leq \max(\tau, 1-\tau)\mathbb{E}(\xi^2|\gamma|^3) \leq m_3\{\tau(1-\tau)+O(h^2)\}\|\mathbf{J}_h^{-1}\boldsymbol{a}\|_{\boldsymbol{\Sigma}}^3.$$

Substituting these bounds into (B.42) yields

$$\sup_{x\in\mathbb{R}}\left|\mathbb{P}\left\{S_n^0 \leq \mathrm{var}(S_n)^{1/2}x\right\} - \Phi(x)\right| \leq c_2 n^{-1/2} \tag{B.43}$$

for some constant $c_2 > 0$.

Recall that $\sigma_h^2 = \mathbb{E}\{\mathscr{K}_h(-\varepsilon)-\tau\}^2\langle\mathbf{J}_h^{-1}\boldsymbol{a},\boldsymbol{x}\rangle^2 = \mathbb{E}(\gamma\xi)^2$, and thus $|\mathrm{var}(S_n) - \sigma_h^2| = (\mathbb{E}\gamma\xi)^2 \lesssim \|\mathbf{J}_h^{-1}\boldsymbol{a}\|_{\boldsymbol{\Sigma}}^2\cdot h^4$. By an application of Lemma A.7 in the supplement of Spokoiny and Zhilova [2015], for sufficiently small h, we have

$$\sup_{x\in\mathbb{R}}\left|\Phi(x/\mathrm{var}(S_n)^{1/2}) - \Phi(x/\sigma_h)\right| \leq c_3 h^4. \tag{B.44}$$

Before proceeding, we note that the constants c_1–c_3 appeared above are all independent of $\boldsymbol{a}$ and (n,p). Let $G \sim N(0,1)$. Putting together the above derivations, for any $x \in \mathbb{R}$ and $\boldsymbol{a} \in \mathbb{R}^p$, we obtain

$$
\begin{aligned}
&\mathbb{P}\big(n^{1/2}\langle \boldsymbol{a}, \widehat{\boldsymbol{\beta}}_h - \boldsymbol{\beta}^*\rangle \le x\big) \\
&\le \mathbb{P}\big\{S_n^0 \le x + c_1 \|\mathbf{J}_h^{-1}\boldsymbol{a}\|_{\boldsymbol{\Sigma}} \cdot n^{1/2}\big(h^{-1/2}\eta_n + h^2\big)\big\} + 3n^{-1} \\
&\le \mathbb{P}\big\{\mathrm{var}(S_n)^{1/2} G \le x + c_1 \|\mathbf{J}_h^{-1}\boldsymbol{a}\|_{\boldsymbol{\Sigma}} \cdot n^{1/2}\big(h^{-1/2}\eta_n + h^2\big)\big\} + c_2 n^{-1/2} + 3n^{-1} \\
&\le \mathbb{P}\big\{\sigma_h G \le x + c_1 \|\mathbf{J}_h^{-1}\boldsymbol{a}\|_{\boldsymbol{\Sigma}} \cdot n^{1/2}\big(h^{-1/2}\eta_n + h^2\big)\big\} + c_2 n^{-1/2} + c_3 h^4 + 3n^{-1} \\
&\le \mathbb{P}\big(\sigma_h G \le x\big) + c_1 (2\pi)^{-1/2} \|\mathbf{J}_h^{-1}\boldsymbol{a}\|_{\boldsymbol{\Sigma}} \cdot n^{1/2}\big(h^{-1/2}\eta_n + h^2\big)/\sigma_h + c_2 n^{-1/2} + c_3 h^4 + 3n^{-1},
\end{aligned}
$$

where the first, second, and third inequalities holds by (B.41), (B.43), and (B.44), respectively, and the last inequality follows from the fact that for any $a \le b$ and $\sigma > 0$, $\Phi(b/\sigma) - \Phi(a/\sigma) \le (2\pi)^{-1/2}(b-a)/\sigma$. Recall that $\sigma_h^2 = \{\tau(1-\tau) + O(h)\}\|\mathbf{J}_h^{-1}\boldsymbol{a}\|_{\boldsymbol{\Sigma}}^2$, $\|\mathbf{J}_h^{-1}\boldsymbol{a}\|_{\boldsymbol{\Sigma}}/\sigma_h \lesssim \{\tau(1-\tau)\}^{-1/2}$ for all sufficiently small h. A similar argument leads to a series of reverse inequalities. The above bounds are independent of x and $\boldsymbol{a}$, and therefore hold uniformly over all x and $\boldsymbol{a}$.

Putting together the pieces, we conclude that under the requirement $\eta_n^{1/2} \lesssim h \lesssim 1$,

$$
\sup_{x \in \mathbb{R}, \boldsymbol{a} \in \mathbb{R}^p} \big|\mathbb{P}\big(n^{1/2}\langle \boldsymbol{a}, \widehat{\boldsymbol{\beta}}_h - \boldsymbol{\beta}^*\rangle \le \sigma_h x\big) - \Phi(x)\big| \lesssim \frac{p + \log n}{(nh)^{1/2}} + n^{1/2}h^2,
$$

as claimed.

Under the additional smoothness condition on $f_{\varepsilon|\boldsymbol{x}}(\cdot)$, using a higher-order Taylor series

expansion of $F_{\varepsilon|\boldsymbol{x}}(\cdot)$ gives

$$\begin{aligned}
&\mathbb{E}\big\{\mathscr{K}(-\varepsilon_i/h)|\boldsymbol{x}_i\big\} = \int_{-\infty}^{\infty} K(u)F_{\varepsilon|\boldsymbol{x}}(hu)\,\mathrm{d}u \\
&= \int_{-\infty}^{\infty} K(u)\bigg[F_{\varepsilon|\boldsymbol{x}}(0) + huf_{\varepsilon|\boldsymbol{x}}(0) + \frac{h^2}{2}u^2 f'_{\varepsilon|\boldsymbol{x}}(0) + \frac{h^3}{3!}u^3 f''_{\varepsilon|\boldsymbol{x}}(0) \\
&\qquad\qquad + \frac{h^3}{2!}u^3\int_0^1 (1-w)^2\big\{f''_{\varepsilon|\boldsymbol{x}}(huw) - f''_{\varepsilon|\boldsymbol{x}}(0)\big\}\mathrm{d}w\bigg]\mathrm{d}u \\
&= \tau + \frac{\kappa_2}{2}f'_{\varepsilon|\boldsymbol{x}}(0)h^2 + O(h^4)\kappa_4 l_2(\boldsymbol{x}),
\end{aligned}$$

from which it follows that

$$\begin{aligned}
\bigg|\mathbb{E}(\gamma_i\xi_i) + \frac{\kappa_2}{2}\langle\mathbf{J}_h^{-1}\boldsymbol{a}, \mathbb{E}\{f'_{\varepsilon|\boldsymbol{x}}(0)\boldsymbol{x}_i\}\rangle\cdot h^2\bigg| &\lesssim \kappa_4\mathbb{E}|\langle\mathbf{J}_h^{-1}\boldsymbol{a}, \boldsymbol{x}_i\rangle l_2(\boldsymbol{x})|\cdot h^4 \\
&\lesssim \kappa_4\|\mathbf{J}_n^{-1}\boldsymbol{a}\|_{\boldsymbol{\Sigma}}\big\{\mathbb{E}l_2^2(\boldsymbol{x})\big\}^{1/2}h^4.
\end{aligned}$$

Together, the above bound and (B.40) with $t = \log n$ imply

$$\begin{aligned}
&\bigg|n^{1/2}\bigg\langle\boldsymbol{a}, \widehat{\boldsymbol{\beta}}_h - \boldsymbol{\beta}^* + \frac{\kappa_2}{2}\mathbf{J}_h^{-1}\mathbb{E}\{f'_{\varepsilon|\boldsymbol{x}}(0)\boldsymbol{x}_i\}\bigg\rangle - S_n^0\bigg| \\
&\qquad\qquad \lesssim \|\mathbf{J}_h^{-1}\boldsymbol{a}\|_{\boldsymbol{\Sigma}}\cdot n^{1/2}\left(\frac{p+\log n}{nh^{1/2}} + h^{3/2}\sqrt{\frac{p+\log n}{n}} + h^4\right) \qquad \text{(B.45)}
\end{aligned}$$

with probability at least $1-3n^{-1}$. Repeating the above analysis, with (B.41) replaced by (B.45), proves the refined Berry-Esseen bound (3.26). □

B.2.6 Proof of Theorem 3.3.4

Keep the notation used in the proof of Theorem 3.3.1. With non-negative multipliers w_i's, the weighted loss $\widehat{Q}_h^\flat(\cdot)$ given in (3.8) is also convex. For $\boldsymbol{\delta}\in\mathbb{R}^p$, define the bootstrap counterpart of $\widehat{D}_h(\cdot)$ as $\widehat{D}_h^\flat(\boldsymbol{\delta}) = \widehat{Q}_h^\flat(\boldsymbol{\beta}^*+\boldsymbol{\delta}) - \widehat{Q}_h^\flat(\boldsymbol{\beta}^*)$. It is easy to see that $\mathbb{E}^*\{\widehat{D}_h^\flat(\boldsymbol{\delta})\} = \widehat{D}_h(\boldsymbol{\delta})$. Similarly

to (B.12), we have

$$\begin{aligned}\widehat{D}_h^\flat(\boldsymbol{\delta}) &= \widehat{D}_h(\boldsymbol{\delta}) + \{\widehat{D}_h^\flat(\boldsymbol{\delta}) - \widehat{D}_h(\boldsymbol{\delta})\} \\ &\geq R_h(\boldsymbol{\delta}) - 0.5 l_0 \kappa_2 h^2 \cdot \|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}} - \{D_h(\boldsymbol{\delta}) - \widehat{D}_h(\boldsymbol{\delta})\} - \{\widehat{D}_h(\boldsymbol{\delta}) - \widehat{D}_h^\flat(\boldsymbol{\delta})\} \\ &\geq \frac{1}{2}\big(\underline{f} - l_0\kappa_1 h - 0.5 l_0 m_3 \|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}} - l_0\kappa_2 h^2/\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}}\big)\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}}^2 \\ &\qquad - \underbrace{\{D_h(\boldsymbol{\delta}) - \widehat{D}_h(\boldsymbol{\delta})\}}_{\text{sampling error}} - \underbrace{\{\widehat{D}_h(\boldsymbol{\delta}) - \widehat{D}_h^\flat(\boldsymbol{\delta})\}}_{\text{bootstrap error}}, \quad \boldsymbol{\delta} \in \mathbb{R}^p. \end{aligned} \tag{B.46}$$

As before, let $r_0 = (2\kappa_2/m_3)^{1/2}h$ be an intermediate convergence radius, and set $r_l = r_0 h$. For any $\boldsymbol{\delta} \in \partial\Theta(r_0)$, it follows that

$$\widehat{D}_h^\flat(\boldsymbol{\delta}) \geq \frac{1}{2}\big\{\underline{f} - l_0\kappa_1 h - l_0(2\kappa_2/m_3)^{1/2}h\big\}r_0^2 - \{D_h(\boldsymbol{\delta}) - \widehat{D}_h(\boldsymbol{\delta})\} - \{\widehat{D}_h(\boldsymbol{\delta}) - \widehat{D}_h^\flat(\boldsymbol{\delta})\}. \tag{B.47}$$

As a bootstrap counterpart of Lemma B.2.4, the following lemma provides high probability bounds for the bootstrap error $\widehat{D}_h(\boldsymbol{\delta}) - \widehat{D}_h^\flat(\boldsymbol{\delta})$ uniformly over $\boldsymbol{\delta} \in \mathbb{R}^p$ in some compact subset.

Lemma B.2.4. *For each $t \geq 0$, there exists an event $\mathscr{E}_1(t)$ with $\mathbb{P}\{\mathscr{E}_1(t)\} \geq 1 - e^{-t}$ such that conditioned on $\mathscr{E}_1(t)$,*

$$\mathbb{P}^*\left\{\sup_{\boldsymbol{\delta}\in\Theta(r)} \{\widehat{D}_h(\boldsymbol{\delta}) - \widehat{D}_h^\flat(\boldsymbol{\delta})\} \geq C\bar{\tau}\upsilon_1 r\sqrt{\frac{u}{n}}\right\} \leq e^{2p-u}$$

for any $u \geq 0$ as long as $n \gtrsim p + t$. Moreover, for any $r_u > r_l > 0$, with $\mathbb{P}^*$-probability (over $\{e_i\}_{i=1}^n$) at least $1 - \lceil e\log(\frac{r_u}{r_l})\rceil e^{2p-u}$ conditioned on $\mathscr{E}_1(t)$,

$$\widehat{D}_h(\boldsymbol{\delta}) - \widehat{D}_h^\flat(\boldsymbol{\delta}) \leq C'\bar{\tau}\upsilon_1\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}}\sqrt{\frac{u}{n}}$$

holds for all $\boldsymbol{\delta} \in \mathbb{R}^p$ satisfying $r_l \leq \|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}} \leq r_u$ as long as $n \gtrsim p + t$. Here both $C, C' > 0$ are

absolute constants.

Let $\mathscr{E}_1(t)$ be the event in Lemma B.2.4 that occurs with probability at least $1-e^{-t}$. Applying the first inequality with $r = r_0$ and $u = 2p+t$, we obtain that conditioned on $\mathscr{E}_1(t)$,

$$\widehat{D}_h(\boldsymbol{\delta}) - \widehat{D}_h^\flat(\boldsymbol{\delta}) \le C\bar{\tau}\upsilon_1 r_0 \sqrt{\frac{2p+t}{n}}.$$

Let $\mathscr{E}_2(t)$ be the event that the bounds (B.17) and (B.18) hold. Then, the "good event" $\mathscr{E}(t) := \mathscr{E}_1(t) \cap \mathscr{E}_2(t)$ satisfies $\mathbb{P}\{\mathscr{E}(t)\} \ge 1-3e^{-t}$. By (B.47) and the above bound, we find that with $\mathbb{P}^*$-probability at least $1-e^{-t}$ conditioned on $\mathscr{E}(t)$, $\widehat{D}_h^\flat(\boldsymbol{\delta}) > 0$ for all $\boldsymbol{\delta} \in \partial\Theta(r_0)$ as long as the bandwidth satisfies $\underline{f}^{-1} m_3^{1/2} \upsilon_1 \sqrt{(p+t)/n} \lesssim h \lesssim \underline{f} m_3^{-1/2}$. Let $\widehat{\boldsymbol{\delta}}^\flat = \widehat{\boldsymbol{\beta}}_h^\flat - \boldsymbol{\beta}^*$. Then, $\widehat{D}_h^\flat(\widehat{\boldsymbol{\delta}}^\flat) \le 0$ by the optimality of $\widehat{\boldsymbol{\beta}}_h^\flat$. This, combined with the convexity of $\widehat{Q}_h^\flat(\cdot)$, ensures that $\widehat{\boldsymbol{\delta}}^\flat \in \Theta(r_0)$.

Next, consider the decomposition $\Theta(r_0) = \Theta(r_l) \cap \Theta(r_l, r_0)$, where $\Theta(r_l, r_0) = \{\boldsymbol{\delta} \in \mathbb{R}^d : r_l \le \|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}} \le r_0\}$ and $r_l = r_0 h = (2\kappa_2/m_3)^{1/2} h^2$. If $\widehat{\boldsymbol{\delta}}^\flat \in \Theta(r_l)$, the claimed bound holds trivially. Throughout the rest, we assume $\widehat{\boldsymbol{\delta}}^\flat \in \Theta(r_l, r_0)$. Taking $u = \log(e\log h^{-1}) + 2p + t$ in the second inequality in Lemma B.2.4 yields that, with $\mathbb{P}^*$-probability at least $1-e^{-t}$ conditioned on $\mathscr{E}(t)$,

$$\widehat{D}_h(\widehat{\boldsymbol{\delta}}^\flat) - \widehat{D}_h^\flat(\widehat{\boldsymbol{\delta}}^\flat) \le \|\widehat{\boldsymbol{\delta}}^\flat\|_{\boldsymbol{\Sigma}} \cdot \underbrace{C'\bar{\tau}\upsilon_1 \sqrt{\frac{\log(e\log h^{-1}) + 2p + t}{n}}}_{=: r_1^\flat}.$$

Substituting the above bound and (B.18) into (B.46), we conclude that

$$(\underline{f} - l_0\kappa_1 h)\|\widehat{\boldsymbol{\delta}}^\flat\|_{\boldsymbol{\Sigma}}^2 \le (2r_1 + 2r_1^\flat + l_0\kappa_2 h^2)\|\widehat{\boldsymbol{\delta}}^\flat\|_{\boldsymbol{\Sigma}} + 0.5 l_0 m_3 \|\widehat{\boldsymbol{\delta}}^\flat\|_{\boldsymbol{\Sigma}}^3 \le 2(r_1 + r_1^\flat + l_0\kappa_2 h^2)\|\widehat{\boldsymbol{\delta}}^\flat\|_{\boldsymbol{\Sigma}}.$$

Canceling out a factor of $\|\widehat{\boldsymbol{\delta}}^\flat\|_{\boldsymbol{\Sigma}}$ from both sides yields the claimed bound. □

Proof of Lemma B.2.4

We will use a similar argument as in the proof of Lemma B.2.1. Consider the bootstrap loss difference $\widehat{D}_h^\flat(\boldsymbol{\delta}) - \widehat{D}_h(\boldsymbol{\delta}) = (1/n)\sum_{i=1}^n e_i d_h(\boldsymbol{\delta}; \boldsymbol{z}_i)$, where $d_h(\boldsymbol{\delta}; \boldsymbol{z}_i) = \ell_h(\varepsilon_i - \langle \boldsymbol{x}_i, \boldsymbol{\delta}\rangle) -$

$\ell_h(\varepsilon_i)$ with $\boldsymbol{z}_i = (\boldsymbol{x}_i, \varepsilon_i)$, and e_i's are independent Rademacher random variables that are independent of $\{\boldsymbol{z}_i\}_{i=1}^n$.

For any $r > 0$ and any $\varepsilon \in (0,1)$, applying a conditional version of Chernoff's inequality to $\Delta_\varepsilon^\flat(r) := n(1-\varepsilon)\sup_{\boldsymbol{\delta}\in\Theta(r)}\{\widehat{D}_h(\boldsymbol{\delta}) - \widehat{D}_h^\flat(\boldsymbol{\delta})\}/(\bar{\tau}r)$ yields

$$\mathbb{P}^*\big\{\Delta_\varepsilon^\flat(r) \geq u\big\} \leq \exp\left[-\sup_{\lambda\geq 0}\big\{\lambda u - \log \mathbb{E}^* e^{\lambda\Delta_\varepsilon^\flat(r)}\big\}\right].$$

Again, by the Ledoux-Talagrand contraction principle and discretization via an ε-net,

$$\begin{aligned}
\mathbb{E}^* e^{\lambda\Delta_\varepsilon^\flat(r)} &\leq \mathbb{E}^* \exp\left\{\frac{\lambda}{r}(1-\varepsilon)\sup_{\boldsymbol{\delta}\in\Theta(r)}\sum_{i=1}^n e_i\langle \boldsymbol{x}_i, \boldsymbol{\delta}\rangle\right\} \\
&\leq \mathbb{E}^* \exp\left\{\lambda(1-\varepsilon)\left\|\sum_{i=1}^n e_i \boldsymbol{w}_i\right\|_2\right\} \leq \mathbb{E}^* \exp\left\{\lambda \max_{1\leq j\leq N_\varepsilon}\sum_{i=1}^n e_i \boldsymbol{u}_j^{\mathrm{T}}\boldsymbol{w}_i\right\} \\
&\leq \sum_{j=1}^{N_\varepsilon}\prod_{i=1}^n \mathbb{E}^* e^{\lambda e_i \boldsymbol{u}_j^{\mathrm{T}}\boldsymbol{w}_i} \leq \sum_{j=1}^{N_\varepsilon} e^{(\lambda^2/2)\sum_{i=1}^n\langle \boldsymbol{u}_j, \boldsymbol{w}_i\rangle^2},
\end{aligned}$$

where we used the Rademacher moment bound $\mathbb{E}e^{\lambda e_i} \leq e^{\lambda^2/2}$ for any $\lambda \in \mathbb{R}$. Here $\{\boldsymbol{u}_j\}_{j=1}^{N_\varepsilon}$ are unit vectors, and $N_\varepsilon \leq (1+2/\varepsilon)^p$. Consequently,

$$\log \mathbb{E}^* e^{\lambda\Delta_\varepsilon^\flat(r)} \leq \log N_\varepsilon + \frac{\lambda^2}{2}\max_{1\leq j\leq N_\varepsilon}\sum_{i=1}^n \langle \boldsymbol{u}_j, \boldsymbol{w}_i\rangle^2,$$

so that for any $u \geq 0$,

$$\begin{aligned}
\sup_{\lambda\geq 0}\big\{\lambda u - \log \mathbb{E}^* e^{\lambda\Delta_\varepsilon^\flat(r)}\big\} &\geq -\log N_\varepsilon + \sup_{\lambda\geq 0}\left(\lambda u - \frac{1}{2}\lambda^2 \max_{1\leq j\leq N_\varepsilon}\sum_{i=1}^n \langle \boldsymbol{u}_j, \boldsymbol{w}_i\rangle^2\right) \\
&= -\log N_\varepsilon + \frac{u^2}{2\max_{1\leq j\leq N_\varepsilon}\sum_{i=1}^n\langle \boldsymbol{u}_j, \boldsymbol{w}_i\rangle^2}.
\end{aligned}$$

Substituting this into the earlier Chernoff's inequality, and by a change of variable, we obtain

that for any $u \geq 0$,

$$\sup_{\boldsymbol{\delta}\in\Theta(r)} \widehat{D}_h(\boldsymbol{\delta}) - \widehat{D}_h^\flat(\boldsymbol{\delta}) \leq \frac{\bar{\tau}}{1-\varepsilon}\sqrt{\max_{1\leq j\leq N_\varepsilon}\frac{1}{n}\sum_{i=1}^n \langle \boldsymbol{u}_j, \boldsymbol{w}_i\rangle^2}\cdot r\sqrt{\frac{2u}{n}}$$

holds with $\mathbb{P}^*$-probability (over $\{e_i\}_{i=1}^n$) at least $1 - e^{p\log(1+2/\varepsilon)-u}$.

The above bound holds for any given $r > 0$. Proceed via a peeling argument, we obtain that for any prespecified $\gamma > 1$ and radii $r_u > r_l > 0$,

$$\widehat{D}_h(\boldsymbol{\delta}) - \widehat{D}_h^\flat(\boldsymbol{\delta}) \leq \frac{\gamma\bar{\tau}}{1-\varepsilon}\sqrt{\max_{1\leq j\leq N_\varepsilon}\frac{1}{n}\sum_{i=1}^n \langle \boldsymbol{u}_j, \boldsymbol{w}_i\rangle^2}\cdot \|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}}\sqrt{\frac{2u}{n}} \text{ holds for all } r_l \leq \|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}} \leq r_u \tag{B.48}$$

with $\mathbb{P}^*$-probability (over $\{e_i\}_{i=1}^n$) at least $1 - \lceil \log(\frac{r_u}{r_l})/\log(\gamma)\rceil e^{p\log(1+2/\varepsilon)-u}$.

Next, we bound the data-dependent quantity $\max_{1\leq j\leq N_\varepsilon}(1/n)\sum_{i=1}^n \langle \boldsymbol{u}_j, \boldsymbol{w}_i\rangle^2$ with $N_\varepsilon \leq (1+2/\varepsilon)^p$ and $\boldsymbol{u}_j \in \mathbb{S}^{p-1}$. Note that $\mathbb{E}\langle \boldsymbol{u}_j, \boldsymbol{w}_i\rangle^2 = 1$. By the sub-Gaussianity of $\boldsymbol{w}_i \in \mathbb{R}^p$ ensured by Condition 3.3.5, for integers $k = 2, 3, \ldots$,

$$\begin{aligned}
\mathbb{E}\big(\langle \boldsymbol{u}_j, \boldsymbol{w}_i\rangle^2\big)^k &\leq \upsilon_1^{2k}\cdot 2k\int_0^\infty \mathbb{P}\big(|\langle \boldsymbol{u}_j, \boldsymbol{w}_i\rangle| \geq \upsilon_1 u\big)u^{2k-1}\,\mathrm{d}u \\
&\leq \upsilon_1^{2k}\cdot 4k\int_0^\infty u^{2k-1}e^{-u^2/2}\,\mathrm{d}u \\
&= 2^k\upsilon_1^{2k}\cdot 2k\int_0^\infty v^{k-1}e^{-v}\,\mathrm{d}v = 2^{k+1}k!\cdot\upsilon_1^{2k}.
\end{aligned}$$

In particular, $\mathbb{E}(\langle \boldsymbol{u}, \boldsymbol{w}_i\rangle^4) \leq 16\upsilon_1^4$ and $\mathbb{E}(\langle \boldsymbol{u}, \boldsymbol{w}_i\rangle^2)^k \leq \frac{k!}{2}\cdot(4\upsilon_1^2)^2\cdot(2\upsilon_1^2)^{k-2}$ for $k \geq 3$. With the above calculations, applying Bernstein's inequality (see, e.g., Theorem 2.10 in Boucheron, Lugosi and Massart [2013]), and taking the union bound over $j = 1, \ldots, N_\varepsilon$, we obtain that for any $v \geq 0$,

$$\mathbb{P}\left(\max_{1\leq j\leq N_\varepsilon}\frac{1}{n}\sum_{i=1}^n \langle \boldsymbol{u}_j, \boldsymbol{w}_i\rangle^2 \geq 1 + 4\upsilon_1^2\sqrt{\frac{2v}{n}} + 2\upsilon_1^2\frac{v}{n}\right) \leq \exp\big\{p\log(1+2/\varepsilon) - v\big\}. \tag{B.49}$$

Finally, we take $\varepsilon = 2/(e^2-1)$, $\gamma = e^{1/e}$ and $v = 2p+t$ ($t \geq 0$) in (B.48) and (B.49). Let $\mathscr{E}_1(t)$ be the event that (B.49) holds. Then, $\mathbb{P}\{\mathscr{E}_1(t)\} \geq 1-e^{-t}$, and with $\mathbb{P}^*$-probability (over $\{e_i\}_{i=1}^n$) at least $1 - \lceil e\log(\frac{r_u}{r_l})\rceil e^{2p-u}$ conditioned on $\mathscr{E}_1(t)$,

$$\widehat{D}_h(\boldsymbol{\delta}) - \widehat{D}_h^\flat(\boldsymbol{\delta}) \leq 3\bar{\tau}\sqrt{1+4\upsilon_1^2\sqrt{(4p+2t)/n}+\upsilon_1^2(4p+2t)/n}\cdot\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}}\sqrt{\frac{u}{n}}$$

holds for all $\boldsymbol{\delta} \in \mathbb{R}^p$ satisfying $r_l \leq \|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}} \leq r_u$. This proves the claimed bound under the scaling $n \gtrsim p+t$. □

B.2.7 Proof of Theorem 3.3.5

The proof is based on an argument similar to that used in the proof of Theorem 3.3.2. To begin with, define the random process

$$\Delta^\flat(\boldsymbol{\delta}) = \boldsymbol{\Sigma}^{-1/2}\big\{\nabla\widehat{Q}_h^\flat(\boldsymbol{\beta}^*+\boldsymbol{\delta}) - \nabla\widehat{Q}_h^\flat(\boldsymbol{\beta}^*) - \mathbf{J}_h\boldsymbol{\delta}\big\},\ \ \boldsymbol{\delta}\in\mathbb{R}^p.$$

For a prespecified $r>0$, a key step is to bound the local fluctuation $\sup_{\boldsymbol{\delta}\in\Theta(r)}\|\Delta^\flat(\boldsymbol{\delta})\|_2$. Since $\mathbb{E}(w_i) = \mathbb{E}(1+e_i) = 1$, we have $\mathbb{E}^*\{\nabla\widehat{Q}_h^\flat(\boldsymbol{\beta})\} = \nabla\widehat{Q}_h(\boldsymbol{\beta})$. Define the (conditionally) centered process

$$G^\flat(\boldsymbol{\beta}) = \boldsymbol{\Sigma}^{-1/2}\big\{\nabla\widehat{Q}_h^\flat(\boldsymbol{\beta}) - \nabla\widehat{Q}_h(\boldsymbol{\beta})\big\} = \frac{1}{n}\sum_{i=1}^n\big\{\mathscr{K}_h(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}-y_i)-\tau\big\}e_i\boldsymbol{w}_i,$$

so that $\Delta^\flat(\boldsymbol{\delta})$ be be written as

$$\Delta^\flat(\boldsymbol{\delta}) = \underbrace{\big\{G^\flat(\boldsymbol{\beta}^*+\boldsymbol{\delta}) - G^\flat(\boldsymbol{\beta}^*)\big\}}_{\text{bootstrap error}} + \underbrace{\Delta(\boldsymbol{\delta})}_{\text{sampling error}},$$

where $\Delta(\boldsymbol{\delta})$ is given in (B.36). By the triangle inequality,

$$\sup_{\boldsymbol{\delta}\in\Theta(r)} \|\Delta^\flat(\boldsymbol{\delta})\|_2 \le \sup_{\boldsymbol{\delta}\in\Theta(r)} \|G^\flat(\boldsymbol{\beta}^*+\boldsymbol{\delta}) - G^\flat(\boldsymbol{\beta}^*)\|_2 + \sup_{\boldsymbol{\delta}\in\Theta(r)} \|\Delta(\boldsymbol{\delta})\|_2. \tag{B.50}$$

Let $\mathscr{E}_3(t)$ denote the event that the bound (B.40) holds. It suffices to deal with the first term on the right-hand side of (B.50). Using a change of variable $\boldsymbol{v} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\delta} \in \mathbb{B}^p(r)$ for $\boldsymbol{\delta} \in \Theta(r)$, we have $y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} = \varepsilon_i - \boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{v}$ and

$$\sup_{\boldsymbol{\delta}\in\Theta(r)} \|G^\flat(\boldsymbol{\beta}^*+\boldsymbol{\delta}) - G^\flat(\boldsymbol{\beta}^*)\|_2 \le \sup_{\boldsymbol{v}\in\mathbb{B}^p(r)} \| \underbrace{G^\flat(\boldsymbol{\beta}^*+\boldsymbol{\Sigma}^{-1/2}\boldsymbol{v}) - G^\flat(\boldsymbol{\beta}^*)}_{=:\Delta_0^\flat(\boldsymbol{v})} \|_2, \tag{B.51}$$

where $\Delta_0^\flat(\boldsymbol{v}) = (1/n)\sum_{i=1}^n e_i\boldsymbol{w}_i\{\mathscr{K}_h(\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{v}-\varepsilon_i) - \mathscr{K}_h(-\varepsilon_i)\}$ satisfies $\Delta_0^\flat(\mathbf{0}) = \mathbf{0}$ and $\mathbb{E}^*\{\Delta_0^\flat(\boldsymbol{v})\} = \mathbf{0}$. Note that $\nabla\Delta_0^\flat(\boldsymbol{v}) = (1/n)\sum_{i=1}^n e_i K_{i,\boldsymbol{v}}\boldsymbol{w}_i\boldsymbol{w}_i^{\mathrm{T}}$, where $K_{i,\boldsymbol{v}} = K_h(\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{v}-\varepsilon_i)$. For any $\lambda \in \mathbb{R}$ and $\boldsymbol{u}, \boldsymbol{u}' \in \mathbb{S}^{p-1}$, we have

$$\begin{aligned}
&\mathbb{E}^*\exp\{\lambda n^{1/2}\boldsymbol{u}^{\mathrm{T}}\Delta_0^\flat(\boldsymbol{v})\boldsymbol{u}'\} = \prod_{i=1}^n \mathbb{E}^*\exp\{\lambda n^{-1/2}e_i K_{i,\boldsymbol{v}}\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u}\cdot\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u}'\} \\
&\le \prod_{i=1}^n \exp\left\{\frac{\lambda^2}{2n}K_{i,\boldsymbol{v}}^2(\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u}\cdot\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u}')^2\right\} = \exp\left\{\frac{\lambda^2}{2n}\sum_{i=1}^n K_{i,\boldsymbol{v}}^2(\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u}\cdot\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u}')^2\right\}.
\end{aligned}$$

Note that $K_{i,\boldsymbol{v}} \le \kappa_u h^{-1}$, and by Hölder's inequality,

$$\frac{1}{n}\sum_{i=1}^n K_{i,\boldsymbol{v}}^2(\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u}\cdot\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u}')^2 \le \frac{\kappa_u}{h}\cdot\left\{\frac{1}{n}\sum_{i=1}^n K_{i,\boldsymbol{v}}(\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u})^4\right\}^{1/2}\left\{\frac{1}{n}\sum_{i=1}^n K_{i,\boldsymbol{v}}(\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u}')^4\right\}^{1/2}.$$

Define the function $\Lambda_{r,h}(\cdot,\cdot) : \mathbb{R}^p \times \mathbb{R}^p \mapsto [0,\infty)$

$$\Lambda_{r,h}(\boldsymbol{u},\boldsymbol{v}) = \frac{1}{n}\sum_{i=1}^n K_h(r\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{v}-\varepsilon_i)(\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u})^4 \ \text{ for } \boldsymbol{u},\boldsymbol{v}\in\mathbb{S}^{p-1}, \tag{B.52}$$

and write $\|\Lambda_{r,h}\|_\infty = \sup_{\boldsymbol{u},\boldsymbol{v}\in\mathbb{S}^{p-1}}\Lambda_{r,h}(\boldsymbol{u},\boldsymbol{v})$. With this notation, it follows that for any $\lambda \in \mathbb{R}$ and

$\boldsymbol{u},\boldsymbol{v}\in\mathbb{S}^{p-1}$,

$$\sup_{\boldsymbol{\delta}\in\mathbb{B}^p(r)}\mathbb{E}^*\exp\{\lambda n^{1/2}\boldsymbol{u}^{\mathrm{T}}\Delta_0^\flat(\boldsymbol{\delta})\boldsymbol{v}\}\le\exp\left\{\frac{\lambda^2}{2}\kappa_u h^{-1}\|\Lambda_{r,h}\|_\infty\right\}.$$

Thus, applying a conditional version of Theorem A.3 in Spokoiny [2013] yields

$$\sup_{\boldsymbol{v}\in\mathbb{B}^p(r)}\|\Delta_0^\flat(\boldsymbol{v})\|_2\le 6\kappa_u^{1/2}\|\Lambda_{r,h}\|_\infty^{1/2}\cdot r\sqrt{\frac{4p+2t}{nh}} \tag{B.53}$$

with $\mathbb{P}^*$-probability at least $1-e^{-t}$.

Next, we bound the data-dependent quantity $\|\Lambda_{r,h}\|_\infty$. For any $\varepsilon_1,\varepsilon_2\in(0,1)$, there exist ε_1- and ε_2-nets $\{\boldsymbol{u}_1,\ldots,\boldsymbol{u}_{d_1}\}$ and $\{\boldsymbol{v}_1,\ldots,\boldsymbol{v}_{d_2}\}$ of $\mathbb{S}^{p-1}$ with $d_1\le(1+2/\varepsilon_1)^p$ and $d_2\le(1+2/\varepsilon_2)^p$. Given $\boldsymbol{u},\boldsymbol{v}\in\mathbb{S}^{p-1}$, there exist some $1\le j\le d_1$ and $1\le k\le d_2$ such that $\|\boldsymbol{u}-\boldsymbol{u}_j\|_2\le\varepsilon_1$ and $\|\boldsymbol{v}-\boldsymbol{v}_k\|_2\le\varepsilon_2$. At $(\boldsymbol{u},\boldsymbol{v})$, consider the decomposition

$$\Lambda_{r,h}(\boldsymbol{u},\boldsymbol{v})=\Lambda_{r,h}(\boldsymbol{u},\boldsymbol{v})-\Lambda_{r,h}(\boldsymbol{u},\boldsymbol{v}_k)+\Lambda_{r,h}(\boldsymbol{u},\boldsymbol{v}_k).$$

For $\Lambda_{r,h}(\boldsymbol{u},\boldsymbol{v})-\Lambda_{r,h}(\boldsymbol{u},\boldsymbol{v}_k)$, the Lipschitz continuity of $K(\cdot)$ ensures that

$$\begin{aligned}&|\Lambda_{r,h}(\boldsymbol{u},\boldsymbol{v})-\Lambda_{r,h}(\boldsymbol{u},\boldsymbol{v}_k)|\\&\le\frac{l_K r}{nh^2}\sum_{i=1}^n|\boldsymbol{w}_i^{\mathrm{T}}(\boldsymbol{v}-\boldsymbol{v}_k)|(\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u})^4\le\frac{l_K r\varepsilon_2}{h^2}\cdot\max_{1\le i\le n}\|\boldsymbol{w}_i\|_2\cdot\frac{1}{n}\sum_{i=1}^n(\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u})^4.\end{aligned} \tag{B.54}$$

For $\Lambda_{r,h}(\boldsymbol{u},\boldsymbol{v}_k)$, by the triangle inequality for the ℓ_4-norm we have

$$\begin{aligned}\Lambda_{r,h}^{1/4}(\boldsymbol{u},\boldsymbol{v}_k)&=\left\{\frac{1}{n}\sum_{i=1}^n K_h(r\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{v}_k-\varepsilon_i)(\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u})^4\right\}^{1/4}\\&\le\left\{\frac{1}{n}\sum_{i=1}^n K_h(r\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{v}_k-\varepsilon_i)(\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u}_j)^4\right\}^{1/4}+\left\{\frac{1}{n}\sum_{i=1}^n K_h(r\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{v}_k-\varepsilon_i)\langle\boldsymbol{w}_i,\boldsymbol{u}-\boldsymbol{u}_j\rangle^4\right\}^{1/4}\\&\le\Lambda_{r,h}^{1/4}(\boldsymbol{u}_j,\boldsymbol{v}_k)+\varepsilon_1\cdot\sup_{\boldsymbol{u}\in\mathbb{S}^{p-1}}\Lambda_{r,h}^{1/4}(\boldsymbol{u},\boldsymbol{v}_k),\end{aligned}$$

which in turn implies

$$\sup_{\boldsymbol{u}\in\mathbb{S}^{p-1}} \Lambda_{r,h}(\boldsymbol{u},\boldsymbol{v}_k) \le (1-\varepsilon_1)^{-4} \max_{1\le j\le d_1} \Lambda_{r,h}(\boldsymbol{u}_j,\boldsymbol{v}_k). \tag{B.55}$$

In view of (B.54) and (B.55), it suffices to bound $\max_{1\le i\le n}\|\boldsymbol{w}_i\|_2$, $\sup_{\boldsymbol{u}\in\mathbb{S}^{p-1}}(1/n)\sum_{i=1}^n(\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u})^4$ and $\max_{(j,k)\in[d_1]\times[d_2]}\Lambda_{r,h}(\boldsymbol{u}_j,\boldsymbol{v}_k)$.

Lemma B.2.5. *For any $t\ge 0$, $\max_{1\le i\le n}\|\boldsymbol{w}_i\|_2^2 \le C_1\upsilon_1^2(p+\log n+t)$ with probability at least $1-e^{-t}$, where $C_1>0$ is an absolute constant.*

For the supremum $\sup_{\boldsymbol{u}\in\mathbb{S}^{p-1}}(1/n)\sum_{i=1}^n(\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u})^4$, similarly to (B.55) it can be shown that

$$\sup_{\boldsymbol{u}\in\mathbb{S}^{p-1}} \frac{1}{n}\sum_{i=1}^n(\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u})^4 \le (1-\varepsilon_1)^{-4}\max_{1\le j\le d_1}\frac{1}{n}\sum_{i=1}^n(\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u}_j)^4.$$

Fix j and k, Condition 3.3.5 implies

$$\begin{aligned}\mathbb{E}e^{\{(\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u}_j)^4/(36\upsilon_1^4)\}^{1/2}} &= \mathbb{E}e^{(\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u}_j)^2/(6\upsilon_1^2)} = 1+\int_0^\infty e^u\mathbb{P}\big\{|\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u}_j|\ge\upsilon_1(6u)^{1/2}\big\}\mathrm{d}u\\ &\le 1+2\int_0^\infty e^{u-3u}\mathrm{d}u = 1+1=2.\end{aligned}$$

Therefore, $\|(\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u}_j)^4\|_{\psi_{1/2}}\le 36\upsilon_1^4$, where $\|\cdot\|_{\psi_r}$ denotes the ψ_r-norm ($r>0$); see Definition 2.1 in Adamczak et al. [2011]. Since $0\le K_h(r\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{v}_k-\varepsilon_i)\le\kappa_u h^{-1}$, it is easy to see that $\|K_h(r\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{v}_k-\varepsilon_i)(\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u}_j)^4\|_{\psi_{1/2}}\le 36\kappa_u\upsilon_1^4h^{-1}$. Moreover, note that $\mathbb{E}(\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u}_j)^4\le m_4$ and

$$\mathbb{E}\big\{K_h(r\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{v}_k-\varepsilon_i)(\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u}_j)^4\big\} = \mathbb{E}\big[\mathbb{E}\big\{K_h(r\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{v}_k-\varepsilon_i)|\boldsymbol{x}_i\big\}(\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u}_j)^4\big]\le\bar{f}m_4.$$

Hence, for any $u,v\ge 3$, applying inequality (3.6) (and those above it) with $s=1/2$ in Adamczak et al. [2011] and the union bound, we obtain that

$$\mathbb{P}\left\{\max_{1\le j\le d_1}\frac{1}{n}\sum_{i=1}^n(\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u}_j)^4\ge m_4+C_2\upsilon_1^4\left(\sqrt{\frac{u}{n}}+\frac{u^2}{n}\right)\right\}\le d_1e^{-u}$$

and

$$\mathbb{P}\left\{\max_{(j,k)\in[d_1]\times[d_2]}\Lambda(\boldsymbol{u}_j,\boldsymbol{v}_k)\geq \bar{f}m_4+C_2\kappa_u\upsilon_1^4\left(\sqrt{\frac{v}{nh^2}}+\frac{v^2}{nh}\right)\right\}\leq d_1d_2e^{-v}.$$

Taking $\varepsilon_1=1-2^{-1/4}$, $\varepsilon_2=n^{-2}$, $u=p\log(1+2/\varepsilon_1)+t$ and $v=p\log\{(1+2/\varepsilon_1)(1+2/\varepsilon_2)\}+t$ in the above bounds, it follows that with probability at least $1-2e^{-t}$,

$$\max_{1\leq j\leq d_1}\frac{1}{n}\sum_{i=1}^n(\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u}_j)^4\leq m_4+C_3\upsilon_1^4\left\{\sqrt{\frac{p+t}{n}}+\frac{(p+t)^2}{n}\right\}$$

and

$$\max_{(j,k)\in[d_1]\times[d_2]}\Lambda(\boldsymbol{u}_j,\boldsymbol{v}_k)\leq \bar{f}m_4+C_4\kappa_u\upsilon_1^4\left\{\sqrt{\frac{p\log n+t}{nh^2}}+\frac{(p\log n+t)^2}{nh}\right\}.$$

Denote by $\mathscr{E}_3(t)$ the event that the above two bounds and the bound in Lemma B.2.5 are satisfied. Then $\mathbb{P}\{\mathscr{E}_3(t)\}\geq 1-3e^{-t}$. Conditioned on $\mathscr{E}_3(t)$ with $t=\log n$, we have

$$\begin{aligned}\|\Lambda_{r,h}\|_\infty &\leq 2\max_{(j,k)\in[d_1]\times[d_2]}\Lambda(\boldsymbol{u}_j,\boldsymbol{v}_k)+\frac{2l_Kr}{n^2h^2}\cdot\max\nolimits_{1\leq i\leq n}\|\boldsymbol{w}_i\|_2\cdot\sup_{1\leq j\leq d_1}\frac{1}{n}\sum_{i=1}^n(\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u}_j)^4\\ &\leq 2\bar{f}m_4+C_5\frac{\upsilon_1^4}{h}\left\{\sqrt{\frac{p\log n}{n}}+\frac{(p\log n)^2}{n}\right\}+C_6\frac{\upsilon_1^5r}{(nh)^2}(p+\log n)^{1/2}. \qquad \text{(B.56)}\end{aligned}$$

With the above preparations, we are ready to prove the Bahadur representation for the bootstrap estimate. Let $\mathscr{E}(t)=\mathscr{E}_1(t)\cap\mathscr{E}_2(t)$ be the event from Theorem 3.3.4 and its proof. In the rest of the proof, we take $t=\log n$, and set the bandwidth $h\asymp(q/n)^{2/5}$ with $q=p+\log n$. Recall that $\widehat{\boldsymbol{\delta}}=\widehat{\boldsymbol{\beta}}_h-\boldsymbol{\beta}^*$ and $\widehat{\boldsymbol{\delta}}^\flat=\widehat{\boldsymbol{\beta}}_h^\flat-\boldsymbol{\beta}^*$. Conditioned on $\mathscr{E}_1(t)\cap\mathscr{E}_2(t)$, $\|\widehat{\boldsymbol{\delta}}\|_{\boldsymbol{\Sigma}}\leq r_{\mathrm{est}}\asymp\sqrt{q/n}$, and $\|\widehat{\boldsymbol{\delta}}^\flat\|_{\boldsymbol{\Sigma}}\leq r_{\mathrm{est}}^\flat\asymp\sqrt{q/n}$ with $\mathbb{P}^*$-probability at least $1-2n^{-1}$. Conditioned further on $\mathscr{E}_3(t)$,

$$\|\boldsymbol{\Sigma}^{-1/2}\mathbf{J}_h(\widehat{\boldsymbol{\beta}}_h-\boldsymbol{\beta}^*)+\boldsymbol{\Sigma}^{-1/2}\nabla\widehat{Q}_h(\boldsymbol{\beta}^*)\|_2=\|\Delta(\widehat{\boldsymbol{\delta}})\|_2\leq\sup_{\boldsymbol{\delta}\in\Theta(r_{\mathrm{est}})}\|\Delta(\boldsymbol{\delta})\|_2\lesssim\left(\frac{q}{n}\right)^{4/5},$$

and with $\mathbb{P}^*$-probability at least $1-3n^{-1}$,

$$\|\boldsymbol{\Sigma}^{-1/2}\mathbf{J}_h(\widehat{\boldsymbol{\beta}}_h^\flat-\boldsymbol{\beta}^*)+\boldsymbol{\Sigma}^{-1/2}\nabla\widehat{Q}_h^\flat(\boldsymbol{\beta}^*)\|_2=\|\Delta^\flat(\widehat{\boldsymbol{\delta}}^\flat)\|_2$$
$$\leq \sup_{\boldsymbol{\delta}\in\Theta(r_{\mathrm{est}}^\flat)}\|\Delta^\flat(\boldsymbol{\delta})\|_2 \lesssim \left(\frac{q}{n}\right)^{4/5}\bigvee\left(\frac{q}{n}\right)^{3/5}\left(\frac{p\log n}{n}\right)^{1/4}\bigvee\left(\frac{q}{n}\right)^{3/5}\frac{p\log n}{n^{1/2}}.$$

Together, the above two bounds proves the claimed result. □

Proof of Lemma B.2.5

Note that, after centering, $\boldsymbol{w}_i=(1,\boldsymbol{w}_{i,-}^{\mathrm{T}})^{\mathrm{T}}$, where $\boldsymbol{w}_{i,-}\in\mathbb{R}^{p-1}$ is a zero-mean sub-Gaussian random vector. Under Condition 3.3.5, there exists some constant $\upsilon_2\asymp\upsilon_1$ depending only on υ_1 such that $\mathbb{E}\exp\{\boldsymbol{\alpha}^{\mathrm{T}}(\boldsymbol{w}-\boldsymbol{e}_1)\}\leq\exp(\|\boldsymbol{\alpha}\|_2^2\upsilon_2^2/2)$ for all $\boldsymbol{\alpha}\in\mathbb{R}^p$, where $\boldsymbol{e}_1=(1,0,\ldots,0)^{\mathrm{T}}$. Then, applying Theorem 2.1 in Hsu, Kakade and Zhang [2012] with $\Sigma=A=\mathbf{I}_p$ yields that, for any $u\geq 0$,

$$\|\boldsymbol{w}_i\|_2^2\leq\upsilon_2^2\big(p+2\sqrt{pu}+2u\big)+1+2(u/p)^{1/2}\leq\upsilon_2^2\big(2p+3u\big)+1+2(u/p)^{1/2}$$

holds with probability at least $1-e^{-u}$. Taking the union bound over $i=1,\ldots,n$ and setting $u=\log n+t$ prove the claimed bound. □

B.2.8 Proof of Proposition 3.3.2

Consider the change of variable $\boldsymbol{v}=\boldsymbol{\Sigma}^{1/2}\boldsymbol{\delta}$, so that $\boldsymbol{\delta}\in\Theta(r)$ is equivalent to $\boldsymbol{v}\in\mathbb{B}^p(r)$. Write $\boldsymbol{w}_i=\boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}_i\in\mathbb{R}^p$, which are isotropic random vectors, and define

$$\widehat{\mathbf{H}}_h(\boldsymbol{v})=\boldsymbol{\Sigma}^{-1/2}\widehat{\mathbf{J}}_h(\boldsymbol{\delta})\boldsymbol{\Sigma}^{-1/2}=\frac{1}{n}\sum_{i=1}^n K_h(\varepsilon_i-\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{v})\boldsymbol{w}_i\boldsymbol{w}_i^{\mathrm{T}},\quad \mathbf{H}_h(\boldsymbol{v})=\mathbb{E}\big\{\widehat{\mathbf{H}}_h(\boldsymbol{v})\big\}. \tag{B.57}$$

For any $\varepsilon_1\in(0,r)$, there exists an ε-net $\mathscr{N}_1:=\{\boldsymbol{v}_1,\ldots,\boldsymbol{v}_{d_1}\}\subseteq\mathbb{B}^p(r)$ with $d_1\leq(1+2r/\varepsilon_1)^p$ satisfying that, for every $\boldsymbol{v}\in\mathbb{B}^p(r)$, there exists some $1\leq j\leq d_1$ such that $\|\boldsymbol{v}-\boldsymbol{v}_j\|_2\leq\varepsilon_1$.

Hence,

$$
\begin{aligned}
&\|\widehat{\mathbf{H}}_h(\boldsymbol{v}) - \mathbf{H}_h(\mathbf{0})\|_2 \\
&\le \|\widehat{\mathbf{H}}_h(\boldsymbol{v}) - \widehat{\mathbf{H}}_h(\boldsymbol{v}_j)\|_2 + \|\widehat{\mathbf{H}}_h(\boldsymbol{v}_j) - \mathbf{H}_h(\boldsymbol{v}_j)\|_2 + \|\mathbf{H}_h(\boldsymbol{v}_j) - \mathbf{H}_h(\mathbf{0})\|_2 \\
&=: I_1(\boldsymbol{v}) + I_{2,j} + I_{3,j}.
\end{aligned}
$$

For $I_1(\boldsymbol{v})$, note that $K_h(u) = (1/h)K(u/h)$ is Lipschitz continuous, i.e. $|K_h(u) - K_h(v)| \le l_K h^{-2}|u-v|$ for all $u, v \in \mathbb{R}$. It follows that

$$
\begin{aligned}
I_1(\boldsymbol{v}) &\le \sup_{\boldsymbol{u},\boldsymbol{u}' \in \mathbb{S}^{p-1}} \frac{1}{n}\sum_{i=1}^n |K_h(\varepsilon_i - \boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{v}) - K_h(\varepsilon_i - \boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{v}_j)| \cdot |\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u} \cdot \boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u}'| \\
&\le l_K h^{-2} \sup_{\boldsymbol{u},\boldsymbol{u}' \in \mathbb{S}^{p-1}} \frac{1}{n}\sum_{i=1}^n |\boldsymbol{w}_i^{\mathrm{T}}(\boldsymbol{v} - \boldsymbol{v}_j) \cdot \boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u} \cdot \boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u}'| \\
&\le l_K h^{-2} \varepsilon_1 \underbrace{\sup_{\boldsymbol{u} \in \mathbb{S}^{p-1}} \frac{1}{n}\sum_{i=1}^n |\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u}|^3}_{=:M_{n,3}}.
\end{aligned} \tag{B.58}
$$

Next, we use the standard covering argument to bound $M_{n,3}$. Given $\varepsilon_2 \in (0,1)$, let $\mathscr{N}_2$ be an ε_2-net of the unit sphere $\mathbb{S}^{p-1}$ with $d_2 := |\mathscr{N}_2| \le (1 + 2/\varepsilon_2)^p$ such that for every $\boldsymbol{u} \in \mathbb{S}^{p-1}$, there exists some $\boldsymbol{u}' \in \mathscr{N}_2$ satisfying $\|\boldsymbol{u} - \boldsymbol{u}'\|_2 \le \varepsilon_2$. Define the (standardized) design matrix $\mathbf{W}_n = n^{-1/3}(\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n)^{\mathrm{T}} \in \mathbb{R}^{n \times p}$, so that $M_{n,3} = \sup_{\boldsymbol{u} \in \mathbb{S}^{p-1}} \|\mathbf{W}_n\boldsymbol{u}\|_3^3$. By the triangle inequality,

$$
\begin{aligned}
\|\mathbf{W}_n\boldsymbol{u}\|_3 &\le \|\mathbf{W}_n\boldsymbol{u}'\|_3 + \|\mathbf{W}_n(\boldsymbol{u} - \boldsymbol{u}')\|_3 \\
&= \|\mathbf{W}_n\boldsymbol{u}'\|_3 + \left(\frac{1}{n}\sum_{i=1}^n |\boldsymbol{w}_i^{\mathrm{T}}(\boldsymbol{u} - \boldsymbol{u}')|^3\right)^{1/3} \le \|\mathbf{W}_n\boldsymbol{u}'\|_3 + \varepsilon_2 \cdot M_{n,3}^{1/3}.
\end{aligned}
$$

Taking the maximum over $\boldsymbol{u}' \in \mathscr{N}_2$, and then taking the supremum over $\boldsymbol{u} \in \mathbb{S}^{p-1}$, we arrive at

$$
M_{n,3} \le (1 - \varepsilon_2)^{-3} \cdot N_{n,3} := (1 - \varepsilon_2)^{-3} \cdot \max_{\boldsymbol{u}' \in \mathscr{N}_2} \frac{1}{n}\sum_{i=1}^n |\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u}'|^3. \tag{B.59}
$$

For every $\boldsymbol{u}' \in \mathscr{N}_2$, note that

$$\mathbb{E}e^{\{|\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u}'|^3/(6^{3/2}\upsilon_1^3)\}^{2/3}} = 1 + \int_0^\infty e^u \mathbb{P}\big\{|\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u}'| \geq \upsilon_1 (6u)^{1/2}\big\}\mathrm{d}u \leq 1 + 2\int_0^\infty e^{-2u}\mathrm{d}u = 2,$$

implying $\||\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u}'|^3\|_{\psi_{2/3}} \leq 6^{3/2}\upsilon_1^3$. Hence, by inequality (3.6) in Adamczak et al. [2011] with $s = 2/3$, we obtain that for any $z \geq 3$,

$$\frac{1}{n}\sum_{i=1}^n |\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u}'|^3 \leq \mathbb{E}|\boldsymbol{w}^{\mathrm{T}}\boldsymbol{u}'|^3 + C_1\upsilon_1^3\left(\sqrt{\frac{z}{n}} + \frac{z^{3/2}}{n}\right)$$

with probability at least $1 - e^{-z}$. Taking the union bound over all vectors $\boldsymbol{u}'$ in $\mathscr{N}_2$ yields that, with probability at least $1 - d_2 e^{-z} \geq 1 - e^{p\log(1+2/\varepsilon_2)-z}$,

$$N_{n,3} \leq m_3 + C_1\upsilon_1^3\left(\sqrt{\frac{z}{n}} + \frac{z^{3/2}}{n}\right)$$

where $m_3 = \sup_{\boldsymbol{u}\in\mathbb{S}^{p-1}} \mathbb{E}|\boldsymbol{w}^{\mathrm{T}}\boldsymbol{u}|^3$. Reorganizing the terms, we get

$$N_{n,3} \leq m_3 + C_1\upsilon_1^3\left[\sqrt{\frac{p\log(1+2/\varepsilon_2)+\log 2+t}{n}} + \frac{\{p\log(1+2/\varepsilon_2)+\log 2+t\}^{3/2}}{n}\right] \quad \text{(B.60)}$$

with probability at least $1 - e^{-t}/2$. Taking $\varepsilon_2 = 1/8$ in (B.59) and (B.60) implies

$$M_{n,3} \leq 1.5m_3 + 1.5C_1\upsilon_1^3\left\{\sqrt{\frac{3p+1+t}{n}} + \frac{(3p+1+t)^{3/2}}{n}\right\}.$$

Under the sample size scaling $n \gtrsim p + t$, plugging the above bound into (B.58) yields

$$\sup_{\boldsymbol{v}\in\mathbb{B}^p(r)} I_1(\boldsymbol{v}) \lesssim \upsilon_1^3(p+t)^{1/2}h^{-2}\varepsilon_1 \quad \text{(B.61)}$$

with probability at least $1 - e^{-t}/2$.

Turning to $I_{2,j}$, note that $\widehat{\mathbf{H}}_h(\boldsymbol{v}_j) - \mathbf{H}_h(\boldsymbol{v}_j) = (1/n)\sum_{i=1}^n (1-\mathbb{E})\phi_{ij}\boldsymbol{w}_i\boldsymbol{w}_i^{\mathrm{T}}$, where $\phi_{ij} =$

$K_h(\varepsilon_i - \boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{v}_j)$ satisfy $|\phi_{ij}| \le \kappa_u h^{-1}$ and

$$\mathbb{E}\big(\phi_{ij}^2|\boldsymbol{x}_i\big) = \frac{1}{h^2}\int_{-\infty}^{\infty} K^2\left(\frac{\langle \boldsymbol{w}_i, \boldsymbol{v}\rangle - t}{h}\right) f_{\varepsilon_i|\boldsymbol{x}_i}(t)\,\mathrm{d}t = \frac{1}{h}\int_{-\infty}^{\infty} K^2(u) f_{\varepsilon_i|\boldsymbol{x}_i}(\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{v} - bu)\,\mathrm{d}u \le \bar{f}\kappa_u h^{-1}$$

almost surely. Given $\varepsilon_3 \in (0, 1/2)$, there exits an ε_3-net $\mathscr{N}_3$ of the sphere $\mathbb{S}^{p-1}$ with $|\mathscr{N}_3| \le (1 + 2/\varepsilon_3)^p$ such that $\|\widehat{\mathbf{H}}_h(\boldsymbol{v}_j) - \mathbf{H}_h(\boldsymbol{v}_j)\|_2 \le (1-2\varepsilon_3)^{-1}\max_{\boldsymbol{u}\in\mathscr{N}_3}|\boldsymbol{u}^{\mathrm{T}}\{\widehat{\mathbf{H}}_h(\boldsymbol{v}_j) - \mathbf{H}_h(\boldsymbol{v}_j)\}\boldsymbol{u}|$. Given $\boldsymbol{u} \in \mathscr{N}_3$ and $k = 2, 3, \ldots$, we bound the higher order moments of $\phi_{ij}(\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u})^2$ by

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

In particular, $\mathbb{E}\phi_{ij}^2(\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u})^4 \le (4\upsilon_1^2)^2\bar{f}\kappa_u h^{-1}$, and

$$\mathbb{E}|\phi_{ij}(\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u})^2|^k \le \frac{k!}{2}\cdot(4\upsilon_1^2)^2\bar{f}\kappa_u h^{-1}\cdot(2\upsilon_1^2\kappa_u h^{-1})^{k-2} \ \text{ for } \ k \ge 3.$$

Applying Bernstein's inequality and the union bound, we find that for any $u \ge 0$,

$$\begin{aligned}
&\|\widehat{\mathbf{H}}_h(\boldsymbol{v}_j) - \mathbf{H}_h(\boldsymbol{v}_j)\|_2 \\
&\le \frac{1}{1-2\varepsilon_3}\max_{\boldsymbol{u}\in\mathscr{N}_3}\left|\frac{1}{n}\sum_{i=1}^{n}(1-\mathbb{E})\phi_{ij}(\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u})^2\right| \le \frac{2\upsilon_1^2}{1-2\varepsilon_3}\left(2\sqrt{2\bar{f}\kappa_u\frac{u}{nh}} + \kappa_u\frac{u}{nh}\right)
\end{aligned}$$

with probability at least $1 - 2(1+2/\varepsilon_3)^p e^{-u} = 1 - (1/2)e^{\log(4)+p\log(1+2/\varepsilon_3)-u}$. Setting $\varepsilon_3 = 2/(e^3-1)$ and $u = \log(4) + 3p + v$, it follows that with probability at least $1 - e^{-v}/2$,

$$I_{2,j} \lesssim \upsilon_1^2\left(\sqrt{\frac{p+v}{nh}} + \frac{p+v}{nh}\right).$$

Once again, taking the union bound over $j = 1, \ldots, d_1$ and setting $v = p\log(1+2r/\varepsilon_1) + t$, we

obtain that with probability at least $1 - d_1 e^{-v} \geq 1 - e^{-t}/2$,

$$\max_{1 \leq j \leq d_1} I_{2,j} \lesssim \sqrt{\frac{p \log(3er/\varepsilon_1) + t}{nh}} + \frac{p \log(3er/\varepsilon_1) + t}{nh}. \tag{B.62}$$

For $I_{3,j}$, following the proof of (B.10) it can similarly shown that $I_{3,j} \leq 0.5 l_0 m_3 r$. Combining this with (B.61) and (B.62), and taking $\varepsilon_1 = r/n \in (0, r)$ in the beginning of the proof, we conclude that with probability at least $1 - e^{-t}$,

$$\sup_{\boldsymbol{\delta} \in \Theta(r)} \|\boldsymbol{\Sigma}^{-1/2}\{\widehat{\mathbf{J}}_h(\boldsymbol{\delta}) - \mathbf{J}_h\}\boldsymbol{\Sigma}^{-1/2}\|_2 \lesssim \sqrt{\frac{p \log n + t}{nh}} + \frac{p \log n + t}{nh} + \frac{(p+t)^{1/2} r}{nh^2} + r$$

as long as $n \gtrsim p + t$. This leads to (3.32) under the prescribed bandwidth constraint.

To derive the same bound for $\widehat{\mathbf{V}}_h(\boldsymbol{\delta}) - \mathbf{V}_h$, notice that $u \mapsto \{\mathscr{K}_h(u) - \tau\}^2$ is a $(2\bar{\tau}\kappa_u/h)$-Lipschitz continuous function, where $\bar{\tau} = \max(\tau, 1-\tau)$. Moreover, for every $\boldsymbol{\delta} \in \mathbb{R}^p$, the random variable $\{\mathscr{K}_h(\langle \boldsymbol{x}_i, \boldsymbol{\delta} \rangle - \varepsilon_i) - \tau\}^2$ takes values in $[0, \bar{\tau}^2]$. We can thus apply the same argument to bound $\sup_{\boldsymbol{\delta} \in \Theta(r)} \|\boldsymbol{\Sigma}^{-1/2}\{\widehat{\mathbf{V}}_h(\boldsymbol{\delta}) - \mathbf{V}_h\}\boldsymbol{\Sigma}^{-1/2}\|_2$. □

B.3 Theoretical Properties of One-step Conquer

In this section, we provide theoretical properties of the one-step conquer estimator $\widehat{\boldsymbol{\beta}}$, defined in Section B.1. The key message is that, when higher-order kernels are used (and if the conditional density $f_{\varepsilon|\boldsymbol{x}}(\cdot)$ has enough derivatives), the asymptotic normality of the one-step estimator holds under weaker growth conditions on p. For example, the scaling condition $p = o(n^{3/8})$ that is required for the conquer estimator can be reduced to roughly $p = o(n^{7/16})$ for the one-step conquer estimator using a kernel of order 4.

Condition 1. *Let $G(\cdot)$ be a symmetric kernel of order $\nu > 2$, that is, $\int_{-\infty}^{\infty} u^k G(u)\,\mathrm{d}u = 0$ for $k = 1, \ldots, \nu - 1$ and $\int_{-\infty}^{\infty} u^\nu G(u)\,\mathrm{d}u \neq 0$. Moreover, $g_k = \int_{-\infty}^{\infty} |u^k G(u)|\,\mathrm{d}u < \infty$ for $1 \leq k \leq \nu$, G is uniformly bounded with $g_u = \sup_{u \in \mathbb{R}} |G(u)| < \infty$ and is l_G-Lipschitz continuous for some $l_G > 0$.*

As before, we write $\mathscr{G}_b(u) = \mathscr{G}(u/b)$ and $\mathscr{G}(u) = \int_{-\infty}^{u} G(v)\,\mathrm{d}v$ for $u \in \mathbb{R}$ and $b > 0$. The use of a higher-order kernel does not necessarily reduce bias unless the conditional density $f_{\varepsilon|\boldsymbol{x}}(\cdot)$ of ε given $\boldsymbol{x}$ is sufficiently smooth. Therefore, we further impose the following smoothness conditions on $f_{\varepsilon|\boldsymbol{x}}(\cdot)$.

Condition 2. *Let $\nu \geq 4$ be the integer in Condition 1. The conditional density $f_{\varepsilon|\boldsymbol{x}}(\cdot)$ is $(\nu-1)$-times differentiable, and satisfies $|f_{\varepsilon|\boldsymbol{x}}^{(\nu-2)}(u) - f_{\varepsilon|\boldsymbol{x}}^{(\nu-2)}(0)| \leq l_{\nu-2}|u|$ for all $u \in \mathbb{R}$ almost surely (over the random vector $\boldsymbol{x}$), where $l_{\nu-2} > 0$ is a constant. Also, there exists some constant $C_G > 0$ such that $\int_{-\infty}^{\infty} |u^{\nu-1}G(u)| \cdot \sup_{|t|\leq|u|} |f_{\varepsilon|\boldsymbol{x}}^{(\nu-1)}(t) - f_{\varepsilon|\boldsymbol{x}}^{(\nu-1)}(0)|\,\mathrm{d}u \leq C_G$ almost surely.*

Notably, we have

$$\nabla Q_b^G(\boldsymbol{\beta}) = \mathbb{E}\big\{\mathscr{G}_b\big(\langle \boldsymbol{x}, \boldsymbol{\beta}\rangle - y\big) - \tau\big\}\boldsymbol{x} \quad \text{and} \quad \nabla^2 Q_b^G(\boldsymbol{\beta}) = \mathbb{E}\big\{G_b(y - \langle \boldsymbol{x}, \boldsymbol{\beta}\rangle)\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\big\}, \tag{B.63}$$

representing the population score and Hessian of $Q_b^G(\cdot) = \mathbb{E}\widehat{Q}_b^G(\cdot)$. As $b \to 0$, we expect $\nabla Q_b^G(\boldsymbol{\beta}^*)$ and $\nabla^2 Q_b^G(\boldsymbol{\beta}^*)$ to converge to $\mathbf{0}$ (zero vector in $\mathbb{R}^p$) and $\mathbf{J} = \mathbb{E}\{f_{\varepsilon|\boldsymbol{x}}(0)\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\}$, respectively. The following proposition validates this claim by providing explicit error bounds.

Proposition B.3.1. *Let $b \in (0,1)$ be a bandwidth. Under Conditions 1 and 2, we have*

$$\begin{aligned}
&\|\boldsymbol{\Sigma}^{-1/2}\nabla Q_b^G(\boldsymbol{\beta}^*)\|_2 \leq l_{\nu-2} g_\nu b^\nu/\nu! \quad \textit{and} \\
&\|\boldsymbol{\Sigma}^{-1/2}\nabla^2 Q_b^G(\boldsymbol{\beta}^*)\boldsymbol{\Sigma}^{-1/2} - \mathbf{H}\|_2 \leq C_G b^{\nu-1}/(\nu-1)!,
\end{aligned}$$

where $\mathbf{H} = \boldsymbol{\Sigma}^{-1/2}\mathbf{J}\boldsymbol{\Sigma}^{-1/2} = \mathbb{E}\{f_{\varepsilon|\boldsymbol{x}}(0)\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}}\}$ with $\boldsymbol{w} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}$.

Proposition B.3.2 shows that when a higher-order kernel is used, the bias is significantly reduced in the sense that $\|\nabla Q_b^G(\boldsymbol{\beta}^*)\|_2 = \mathscr{O}(b^\nu)$ and $\|\nabla^2 Q_b^G(\boldsymbol{\beta}^*) - \mathbf{J}\|_2 = \mathscr{O}(b^{\nu-1})$, where $\nu \geq 4$ is an even integer. Notably, even if the kernel G has negative parts, the population Hessian $\nabla^2 Q_b^G(\boldsymbol{\beta}^*)$ preserves the positive definiteness of $\mathbf{J}$ as long as the bandwidth b is sufficiently small.

To construct the one-step conquer estimator, two key quantities are the sample Hessian $\nabla^2 \widehat{Q}_b^G(\cdot)$ and sample gradient $\nabla \widehat{Q}_b^G(\cdot)$, both evaluated at $\overline{\boldsymbol{\beta}}$, a consistent initial estimate. In the next two propositions, we establish uniform convergence results of the Hessian and gradient of the empirical smoothed loss to their population counterparts. As a direct consequence, $\nabla^2 \widehat{Q}_b^G(\overline{\boldsymbol{\beta}})$ is positive definite with high probability, provided that $\overline{\boldsymbol{\beta}}$ is consistent (i.e., in a local vicinity of $\boldsymbol{\beta}^*$). To be more specific, for $r > 0$, we define the local neighborhood

$$\Theta^*(r) = \big\{\boldsymbol{\beta} \in \mathbb{R}^p : \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma}} \le r\big\}. \tag{B.64}$$

Proposition B.3.2. *Conditions 1, 2 and 3.3.5 ensure that, with probability at least* $1 - e^{-t}$,

$$\sup_{\boldsymbol{\beta} \in \Theta^*(r)} \|\boldsymbol{\Sigma}^{-1/2}\{\nabla^2 \widehat{Q}_b^G(\boldsymbol{\beta}) - \nabla^2 Q_b^G(\boldsymbol{\beta})\}\boldsymbol{\Sigma}^{-1/2}\|_2 \lesssim \sqrt{\frac{p\log n + t}{nb}} + \frac{p\log n + t}{nb} + \frac{(p+t)^{1/2} r}{nb^2}$$

as long as $n \gtrsim p + t$.

Proposition B.3.3. *Conditions 1, 2 and 3.3.5 ensure that, with probability at least* $1 - e^{-t}$,

$$\sup_{\boldsymbol{\beta} \in \Theta^*(r)} \|\boldsymbol{\Sigma}^{-1/2}\{\nabla \widehat{Q}_b^G(\boldsymbol{\beta}) - \nabla \widehat{Q}_b^G(\boldsymbol{\beta}^*)\} - \boldsymbol{\Sigma}^{-1/2}\mathbf{J}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)\|_2 \lesssim r\left(\sqrt{\frac{p+t}{nb}} + r + b^{\nu-1}\right) \tag{B.65}$$

as long as $\sqrt{(p+t)/n} \lesssim b$.

With the above preparations, we are ready to present the Bahadur representation for the one-step conquer estimator $\widehat{\boldsymbol{\beta}}$.

Theorem B.3.1. *Assume Conditions 3.3.1, 3.3.2 and 3.3.5 in the main text and Conditions 1 and 2 hold. For any* $t > 0$, *let the sample size* n, *dimension* p *and the bandwidths* $h, b > 0$ *satisfy* $n \gtrsim p(\log n)^2 + t$, $\sqrt{(p+t)/n} \lesssim h \lesssim \{(p+t)/n\}^{1/4}$ *and* $\sqrt{(p+t)/n} \lesssim b \lesssim \{(p+t)/n\}^{1/(2\nu)}$.

Then, the one-step conquer estimator $\widehat{\boldsymbol{\beta}}$ satisfies the bound

$$\begin{aligned}&\left\|\boldsymbol{\Sigma}^{-1/2}\mathbf{J}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)-\frac{1}{n}\sum_{i=1}^{n}\big\{\tau-\mathscr{G}_b(-\varepsilon_i)\big\}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}_i\right\|_2\\&\quad\lesssim\left\{\underbrace{\sqrt{(p\log n+t)/(nb)}}_{\text{variance term}}+\underbrace{b^{\nu-1}}_{\text{bias term}}\right\}\sqrt{\frac{p+t}{n}}\end{aligned}\tag{B.66}$$

with probability at least $1-5e^{-t}$.

Theorem B.3.1 shows that using a higher-order kernel ($\nu\geq 4$) allows one to choose larger bandwidth, thereby reducing the "variance" and the total Bahadur linearization error. Similarly to Theorem 3.3.3 in the main text, the following asymptotic normal approximation result for linear projections of one-step conquer is a direct consequence of Theorem B.3.1.

Theorem B.3.2. *Assume Conditions 3.3.1, 3.3.2 and 3.3.5 in the main text and Conditions 1 and 2 hold. Let the bandwidths satisfy $(q/n)^{1/2}\lesssim h\lesssim(q/n)^{1/4}$ and $(q/n)^{1/2}\lesssim b\lesssim(q/n)^{1/(2\nu)}$, where $q:=p+\log n$. Then,*

$$\sup_{x\in\mathbb{R},\boldsymbol{a}\in\mathbb{R}^p}\big|\mathbb{P}\big(n^{1/2}\langle\boldsymbol{a},\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\rangle/\sigma_0\leq x\big)-\Phi(x)\big|\lesssim\sqrt{\frac{(p+\log n)p\log n}{nb}}+n^{1/2}b^{\nu},\tag{B.67}$$

where $\sigma_0^2=\sigma_0^2(\boldsymbol{a})=\tau(1-\tau)\|\mathbf{J}^{-1}\boldsymbol{a}\|_{\boldsymbol{\Sigma}}^2$. In particular, with $b\asymp(q/n)^{2/(2\nu+1)}$,

$$\sup_{x\in\mathbb{R},\boldsymbol{a}\in\mathbb{R}^p}\big|\mathbb{P}\big(n^{1/2}\langle\boldsymbol{a},\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\rangle\leq\sigma_0 x\big)-\Phi(x)\big|\to 0$$

as $n\to\infty$ under the scaling $p^{4\nu/(2\nu-1)}(\log n)^{(2\nu+1)/(2\nu-1)}=o(n)$.

Let $G(\cdot)$ be a kernel of order $\nu=4$. In view of Theorem B.3.2, we take $h\asymp\{(p+\log n)/n\}^{2/5}$ as in the main text and $b=\{(p+\log n)/n\}^{2/9}$, thereby obtaining that $n^{1/2}\langle\boldsymbol{a},\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\rangle$, for an arbitrary $\boldsymbol{a}\in\mathbb{R}^p$, is asymptotically normally distributed as long as $p(\log n)^{9/16}=o(n^{7/16})$ as $n\to\infty$.

B.3.1 Proof of Proposition B.3.1

We start from the gradient $\boldsymbol{\Sigma}^{-1/2}\nabla Q_b^G(\boldsymbol{\beta}^*) = \mathbb{E}\{\mathscr{G}_b(-\varepsilon)-\tau\}\boldsymbol{w}$ with $\boldsymbol{w} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}$. Let $\mathbb{E}_{\boldsymbol{x}}$ be the conditional expectation given $\boldsymbol{x}$. By integration by parts,

$$\mathbb{E}_{\boldsymbol{x}}\mathscr{G}_b(-\varepsilon) = \int_{-\infty}^{\infty}\mathscr{G}(-t/b)\,\mathrm{d}F_{\varepsilon|\boldsymbol{x}}(t) = \int_{-\infty}^{\infty}G(u)F_{\varepsilon|\boldsymbol{x}}(-bu)\,\mathrm{d}u. \tag{B.68}$$

Applying a Taylor series expansion with integral remainder on $F_{\varepsilon|\boldsymbol{x}}(-bu)$ yields

$$\begin{aligned}
F_{\varepsilon|\boldsymbol{x}}(-bu) &= F_{\varepsilon|\boldsymbol{x}}(0) + \sum_{\ell=1}^{\nu-1} F_{\varepsilon|\boldsymbol{x}}^{(\ell)}(0)\frac{(-bu)^\ell}{\ell!} \\
&\quad + \frac{(-bu)^{\nu-1}}{(\nu-2)!}\int_0^1 (1-w)^{\nu-2}\big\{F_{\varepsilon|\boldsymbol{x}}^{(\nu-1)}(-buw) - F_{\varepsilon|\boldsymbol{x}}^{(\nu-1)}(0)\big\}\mathrm{d}w \\
&= \tau + \sum_{\ell=0}^{\nu-2} f_{\varepsilon|\boldsymbol{x}}^{(\ell)}(0)\frac{(-bu)^{\ell+1}}{(\ell+1)!} \\
&\quad + \frac{(-bu)^{\nu-1}}{(\nu-2)!}\int_0^1 (1-w)^{\nu-2}\big\{f_{\varepsilon|\boldsymbol{x}}^{(\nu-2)}(-buw) - f_{\varepsilon|\boldsymbol{x}}^{(\nu-2)}(0)\big\}\mathrm{d}w.
\end{aligned}$$

Recall that G is a kernel of order $\nu \geq 4$ (an even integer) and $g_\nu = \int_{-\infty}^{\infty}|u^\nu G(u)|\,\mathrm{d}u < \infty$. Substituting the above expansion into (B.68), we obtain

$$\mathbb{E}_{\boldsymbol{x}}\mathscr{G}_b(-\varepsilon) = \tau - \frac{b^{\nu-1}}{(\nu-2)!}\int_{-\infty}^{\infty}\int_0^1 u^{\nu-1}G(u)(1-w)^{\nu-2}\big\{f_{\varepsilon|\boldsymbol{x}}^{(\nu-2)}(-buw) - f_{\varepsilon|\boldsymbol{x}}^{(\nu-2)}(0)\big\}\mathrm{d}w\mathrm{d}u.$$

Furthermore, by the Lipschitz continuity of $f_{\varepsilon|\boldsymbol{x}}^{(\nu-2)}(\cdot)$ around 0,

$$\begin{aligned}
|\mathbb{E}_{\boldsymbol{x}}\mathscr{G}_b(-\varepsilon) - \tau| &\leq \frac{l_{\nu-2}b^\nu}{(\nu-2)!}\int_{-\infty}^{\infty}\int_0^1 |u^\nu G(u)|(1-w)^{\nu-2}w\,\mathrm{d}w\mathrm{d}u \\
&= B(2,\nu-1)l_{\nu-2}g_\nu b^\nu/(\nu-2)!,
\end{aligned}$$

where $B(x,y) := \int_0^1 t^{x-1}(1-t)^{y-1}\mathrm{d}t$ denotes the beta function. In particular, $B(2,\nu-1) = \Gamma(2)\Gamma(\nu-1)/\Gamma(\nu+1) = (\nu-2)!/\nu!$. Putting together the pieces yields

$$\|\boldsymbol{\Sigma}^{-1/2}\nabla Q_b^G(\boldsymbol{\beta}^*)\|_2 = \sup_{\boldsymbol{u}\in\mathbb{S}^{p-1}} \mathbb{E}\mathbb{E}_{\boldsymbol{x}}\big\{\mathscr{G}_b(-\varepsilon)-\tau\big\}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{u} \le l_{\nu-2}g_\nu b^\nu/\nu!.$$

Turning to the Hessian, note that

$$\|\boldsymbol{\Sigma}^{-1/2}\{\nabla^2 Q_b^G(\boldsymbol{\beta}^*)-\mathbf{J}\}\boldsymbol{\Sigma}^{-1/2}\|_2 = \left\|\mathbb{E}\int_{-\infty}^{\infty} G(u)\big\{f_{\varepsilon|\boldsymbol{x}}(-bu)-f_{\varepsilon|\boldsymbol{x}}(0)\big\}\mathrm{d}u\,\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}}\right\|_2.$$

Applying a similar Taylor series expansion as above, we have

$$f_{\varepsilon|\boldsymbol{x}}(t) = f_{\varepsilon|\boldsymbol{x}}(0) + \sum_{\ell=1}^{\nu-1} f_{\varepsilon|\boldsymbol{x}}^{(\ell)}(0)\frac{t^\ell}{\ell!} + \frac{t^{\nu-1}}{(\nu-2)!}\int_0^1 (1-w)^{\nu-2}\big\{f_{\varepsilon|\boldsymbol{x}}^{(\nu-1)}(tw)-f_{\varepsilon|\boldsymbol{x}}^{(\nu-1)}(0)\big\}\mathrm{d}w. \tag{B.69}$$

Under Conditions 1 and 2, it follows that

$$\begin{aligned}
&\|\boldsymbol{\Sigma}^{-1/2}\{\nabla^2 Q_b^G(\boldsymbol{\beta}^*)-\mathbf{J}\}\boldsymbol{\Sigma}^{-1/2}\|_{\boldsymbol{\Omega}} \\
&\le \frac{b^{\nu-1}}{(\nu-2)!}\sup_{\boldsymbol{u},\boldsymbol{\delta}\in\mathbb{S}^{p-1}} \mathbb{E}\int_{-\infty}^{\infty}\int_0^1 u^{\nu-1}G(u)(1-w)^{\nu-2}\times \\
&\qquad\qquad \big\{f_{\varepsilon|\boldsymbol{x}}^{(\nu-1)}(-buw)-f_{\varepsilon|\boldsymbol{x}}^{(\nu-1)}(0)\big\}\mathrm{d}w\,\mathrm{d}u\,\big\langle\boldsymbol{w},\boldsymbol{u}\big\rangle\big\langle\boldsymbol{w},\boldsymbol{\delta}\big\rangle \\
&\le \frac{b^{\nu-1}}{(\nu-1)!}\sup_{\boldsymbol{u},\boldsymbol{\delta}\in\mathbb{S}^{p-1}} \mathbb{E}\int_{-\infty}^{\infty}|u^{\nu-1}G(u)|\sup_{|t|\le b|u|}\big|f_{\varepsilon|\boldsymbol{x}}^{(\nu-1)}(t)-f_{\varepsilon|\boldsymbol{x}}^{(\nu-1)}(0)\big|\,\mathrm{d}u\cdot|\langle\boldsymbol{w},\boldsymbol{u}\rangle\langle\boldsymbol{w},\boldsymbol{\delta}\rangle| \\
&\le \frac{C_G b^{\nu-1}}{(\nu-1)!}\sup_{\boldsymbol{u}\in\mathbb{S}^{p-1}}\mathbb{E}\big\langle\boldsymbol{w},\boldsymbol{u}\big\rangle^2 = \frac{C_G}{(\nu-1)!}b^{\nu-1}.
\end{aligned}$$

This completes the proof. □

B.3.2 Proof of Proposition B.3.2

The proof is almost identical to that of Proposition 3.3.2, and thus is omitted. □

B.3.3 Proof of Proposition B.3.3

Define the stochastic process $\Delta_b(\boldsymbol{\beta}) = \boldsymbol{\Sigma}^{-1/2}\{\nabla \widehat{Q}_b^G(\boldsymbol{\beta}) - \nabla \widehat{Q}_b^G(\boldsymbol{\beta}^*) - \mathbf{J}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)\}$. By the triangle inequality,

$$\sup_{\boldsymbol{\beta}\in\Theta^*(r)} \|\Delta_b(\boldsymbol{\beta})\|_2 \leq \sup_{\boldsymbol{\beta}\in\Theta^*(r)} \|\mathbb{E}\Delta_b(\boldsymbol{\beta})\|_2 + \sup_{\boldsymbol{\beta}\in\Theta^*(r)} \|\Delta_b(\boldsymbol{\beta}) - \mathbb{E}\Delta_b(\boldsymbol{\beta})\|_2$$

Recall that $\mathbf{H} = \boldsymbol{\Sigma}^{-1/2}\mathbf{J}\boldsymbol{\Sigma}^{-1/2} = \mathbb{E}\{f_{\varepsilon|\boldsymbol{x}}(0)\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}}\}$. For the first term on the right-hand side, using the mean value theorem for vector-valued functions yields

$$\mathbb{E}\Delta_b(\boldsymbol{\beta}) = \left\{\boldsymbol{\Sigma}^{-1/2}\int_0^1 \nabla^2 Q_b^G\big((1-s)\boldsymbol{\beta}^* + s\boldsymbol{\beta}\big)\,\mathrm{d}s\,\boldsymbol{\Sigma}^{-1/2} - \mathbf{J}_0\right\}\boldsymbol{\Sigma}^{1/2}(\boldsymbol{\beta} - \boldsymbol{\beta}^*).$$

By a change of variable $\boldsymbol{\delta} = \boldsymbol{\Sigma}^{1/2}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)$,

$$\nabla^2 Q_b^G\big((1-s)\boldsymbol{\beta}^* + s\boldsymbol{\beta}\big) = \mathbb{E}\int_{-\infty}^{\infty} G(u) f_{\varepsilon|\boldsymbol{x}}(s\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\delta} - bu)\,\mathrm{d}u \cdot \boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}.$$

For every $s \in [0,1]$ and $u \in \mathbb{R}$, it ensures from $f_{\varepsilon|\boldsymbol{x}}(\cdot)$ being Lipschitz that

$$|f_{\varepsilon|\boldsymbol{x}}(s\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\delta} - bu) - f_{\varepsilon|\boldsymbol{x}}(-bu)| \leq l_0 s \cdot |\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\delta}|.$$

Moreover, by the Taylor series expansion (B.69),

$$\begin{aligned} f_{\varepsilon|\boldsymbol{x}}(-bu) = f_{\varepsilon|\boldsymbol{x}}(0) &+ \sum_{\ell=1}^{\nu-1} f_{\varepsilon|\boldsymbol{x}}^{(\ell)}(0)\frac{(-bu)^\ell}{\ell!} \\ &+ \frac{(-bu)^{\nu-1}}{(\nu-2)!}\int_0^1 (1-w)^{\nu-2}\big\{f_{\varepsilon|\boldsymbol{x}}^{(\nu-1)}(-buw) - f_{\varepsilon|\boldsymbol{x}}^{(\nu-1)}(0)\big\}\,\mathrm{d}w. \end{aligned}$$

Consequently,

$$
\begin{aligned}
&\left\| \boldsymbol{\Sigma}^{-1/2} \int_0^1 \nabla^2 Q_b^G\big((1-s)\boldsymbol{\beta}^* + s\boldsymbol{\beta}\big)\,\mathrm{d}s\,\boldsymbol{\Sigma}^{-1/2} - \mathbf{H} \right\|_2 \\
&\le \left\| \frac{b^{\nu-1}}{(\nu-2)!}\mathbb{E}\int_{-\infty}^{\infty}\int_0^1 u^{\nu-1}G(u)(1-w)^{\nu-2}\big\{f_{\varepsilon|\boldsymbol{x}}^{(\nu-1)}(-buw) - f_{\varepsilon|\boldsymbol{x}}^{(\nu-1)}(0)\big\}\,\mathrm{d}w\mathrm{d}u\cdot \boldsymbol{w}\boldsymbol{w}^{\mathrm{T}} \right\|_2 \\
&\quad + 0.5l_0 \cdot \big\|\mathbb{E}\,|\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\delta}|\cdot \boldsymbol{w}\boldsymbol{w}^{\mathrm{T}}\big\|_2 \\
&\le \frac{C_G b^{\nu-1}}{(\nu-1)!}\sup_{\boldsymbol{u}\in\mathbb{S}^{p-1}}\mathbb{E}\langle \boldsymbol{w},\boldsymbol{u}\rangle^2 + \frac{l_0}{2}\sup_{\boldsymbol{u}\in\mathbb{S}^{p-1}}\mathbb{E}\big(|\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\delta}|\langle \boldsymbol{w},\boldsymbol{u}\rangle^2\big) \\
&\le \left(\frac{C_G}{(\nu-1)!}b^{\nu-1} + 0.5l_0m_3\|\boldsymbol{\delta}\|_2\right).
\end{aligned}
$$

Taking the supremum over $\boldsymbol{\beta}\in\Theta^*(r)$, or equivalently $\boldsymbol{\delta}\in\mathbb{B}^p(r)$, yields

$$
\sup_{\boldsymbol{\beta}\in\Theta^*(r)}\|\mathbb{E}\Delta_b(\boldsymbol{\beta})\| \le \left(\frac{C_G}{(\nu-1)!}b^{\nu-1} + 0.5l_0m_3r\right)r \lesssim (b^{\nu-1}+r)r.
$$

For the stochastic term $\sup_{\boldsymbol{\beta}\in\Theta^*(r)}\|\Delta_b(\boldsymbol{\beta}) - \mathbb{E}\Delta_b(\boldsymbol{\beta})\|_2$, following the proof of Theorem 3.3.2, it can be similarly shown that with probability at least $1-e^{-t}$,

$$
\sup_{\boldsymbol{\beta}\in\Theta^*(r)}\|\Delta_b(\boldsymbol{\beta}) - \mathbb{E}\Delta_b(\boldsymbol{\beta})\|_2 \lesssim r\sqrt{\frac{p+t}{nb}}
$$

as long as $\sqrt{(p+t)/n}\lesssim b$.

Combining the last two displays completes the proof of (B.65). □

B.3.4 Proof of Theorem B.3.1

STEP 1 (Consistency of the initial estimate). First, note that the consistency of the initial estimator $\overline{\boldsymbol{\beta}}$—namely, $\overline{\boldsymbol{\beta}}$ lies in a local neighborhood of $\boldsymbol{\beta}^*$ with high probability, is a direct consequence of Theorem 3.3.1. Given a non-negative kernel $K(\cdot)$ and for any $t>0$, the initial estimator $\overline{\boldsymbol{\beta}}$

satisfies

$$\|\overline{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma}} \leq r_1 \asymp \sqrt{\frac{p+t}{n}} \tag{B.70}$$

with probability at least $1-2e^{-t}$ as long as $(\frac{p+t}{n})^{1/2} \lesssim h \lesssim (\frac{p+t}{n})^{1/4}$. Let $\mathscr{E}_{\text{init}}(t)$ be the event that (B.70) holds. Provided that the sample Hessian $\nabla^2 \widehat{Q}_b^G(\overline{\boldsymbol{\beta}})$ is invertible, we have

$$\begin{aligned}
\mathbf{J}(\widehat{\boldsymbol{\beta}} - \overline{\boldsymbol{\beta}}) &= -\mathbf{J}\big\{\nabla^2 \widehat{Q}_b^G(\overline{\boldsymbol{\beta}})\big\}^{-1} \nabla \widehat{Q}_b^G(\overline{\boldsymbol{\beta}}) \\
&= -\mathbf{J}\big\{\nabla^2 \widehat{Q}_b^G(\overline{\boldsymbol{\beta}})\big\}^{-1} \boldsymbol{\Sigma}^{1/2} \cdot \boldsymbol{\Sigma}^{-1/2}\big\{\nabla \widehat{Q}_b^G(\overline{\boldsymbol{\beta}}) - \nabla \widehat{Q}_b^G(\boldsymbol{\beta}^*) - \mathbf{J}(\overline{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\big\} \\
&\quad - \mathbf{J}\big\{\nabla^2 \widehat{Q}_b^G(\overline{\boldsymbol{\beta}})\big\}^{-1} \boldsymbol{\Sigma}^{1/2} \cdot \boldsymbol{\Sigma}^{-1/2}\big\{\mathbf{J}(\overline{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) + \nabla \widehat{Q}_b^G(\boldsymbol{\beta}^*)\big\},
\end{aligned}$$

or equivalently,

$$\begin{aligned}
\boldsymbol{\Sigma}^{-1/2}\mathbf{J}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) &= -\mathbf{H}\widehat{\mathbf{H}}_b^{-1} \cdot \boldsymbol{\Sigma}^{-1/2}\big\{\nabla \widehat{Q}_b^G(\overline{\boldsymbol{\beta}}) - \nabla \widehat{Q}_b^G(\boldsymbol{\beta}^*) - \mathbf{J}(\overline{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\big\} \\
&\quad + \big(\mathbf{I}_p - \mathbf{H}\widehat{\mathbf{H}}_b^{-1}\big)\mathbf{H} \cdot \boldsymbol{\Sigma}^{1/2}(\overline{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) - \mathbf{H}\widehat{\mathbf{H}}_b^{-1} \cdot \boldsymbol{\Sigma}^{-1/2}\nabla \widehat{Q}_b^G(\boldsymbol{\beta}^*),
\end{aligned}$$

where $\mathbf{H} = \mathbb{E}\{f_{\varepsilon|\boldsymbol{x}}(0)\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}}\} = \boldsymbol{\Sigma}^{-1/2}\mathbf{J}\boldsymbol{\Sigma}^{-1/2}$ and $\widehat{\mathbf{H}}_b := \boldsymbol{\Sigma}^{-1/2}\nabla^2 \widehat{Q}_b^G(\overline{\boldsymbol{\beta}})\boldsymbol{\Sigma}^{-1/2}$. It follows that

$$\begin{aligned}
&\|\boldsymbol{\Sigma}^{-1/2}\mathbf{J}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) + \boldsymbol{\Sigma}^{-1/2}\nabla \widehat{Q}_b^G(\boldsymbol{\beta}^*)\|_2 \\
&\leq \|(\mathbf{I}_p - \mathbf{H}\widehat{\mathbf{H}}_b^{-1})\mathbf{H}\|_2 \cdot \|\overline{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma}} + \|\mathbf{I}_p - \mathbf{H}\mathbf{H}_b^{-1}\|_2 \cdot \|\nabla \widehat{Q}_b^G(\boldsymbol{\beta}^*)\|_{\boldsymbol{\Omega}} \\
&\quad + \|\mathbf{H}\widehat{\mathbf{H}}_b^{-1}\|_2 \cdot \|\nabla \widehat{Q}_b^G(\overline{\boldsymbol{\beta}}) - \nabla \widehat{Q}_b^G(\boldsymbol{\beta}^*) - \mathbf{J}(\overline{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_{\boldsymbol{\Omega}}.
\end{aligned} \tag{B.71}$$

In view of (B.71), it remains to bound the following three quantities:

$$\|\mathbf{I}_p - \mathbf{H}\widehat{\mathbf{H}}_b^{-1}\|_2, \ \ \|\nabla \widehat{Q}_b^G(\boldsymbol{\beta}^*)\|_{\boldsymbol{\Omega}} \ \text{ and } \ \|\nabla \widehat{Q}_b^G(\overline{\boldsymbol{\beta}}) - \nabla \widehat{Q}_b^G(\boldsymbol{\beta}^*) - \mathbf{J}(\overline{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_{\boldsymbol{\Omega}}.$$

STEP 2 (Consistency of the sample Hessian $\nabla^2 \widehat{Q}_b^G(\overline{\boldsymbol{\beta}})$). Recall that $\widehat{\mathbf{H}}_b = \boldsymbol{\Sigma}^{-1/2}\nabla^2 \widehat{Q}_b^G(\overline{\boldsymbol{\beta}})\boldsymbol{\Sigma}^{-1/2}$

and $\mathbf{H} = \mathbf{\Sigma}^{-1/2}\mathbf{J}\mathbf{\Sigma}^{-1/2}$. By the triangle inequality,

$$\|\widehat{\mathbf{H}}_b - \mathbf{H}\|_2 \leq \|\mathbf{\Sigma}^{-1/2}\{\nabla^2 \widehat{Q}_b^G(\overline{\boldsymbol{\beta}}) - \nabla^2 Q_b^G(\boldsymbol{\beta}^*)\}\mathbf{\Sigma}^{-1/2}\|_2 + \|\mathbf{\Sigma}^{-1/2}\nabla^2 Q_b^G(\boldsymbol{\beta}^*)\mathbf{\Sigma}^{-1/2} - \mathbf{H}\|_2.$$

Let the bandwidth b satisfy $\max\{\frac{p\log n+t}{n}, n^{-1/2}\} \lesssim b \lesssim 1$. Conditioned on $\mathscr{E}_{\text{init}}(t)$, applying Propositions B.3.1 and B.3.2 with $r = r_1$ yields that, with probability $1 - e^{-t}$,

$$\|\widehat{\mathbf{H}}_b - \mathbf{H}\|_2 \leq r_2 \asymp \sqrt{\frac{p\log n + t}{nb}} + b^{\nu-1}.$$

Under Condition 3.3.2, $0 < \underline{f} \leq \lambda_{\min}(\mathbf{H}) \leq \lambda_{\max}(\mathbf{H}) \leq \bar{f}$, so that $\|\mathbf{H}^{-1}\|_2 \leq \underline{f}^{-1}$. For sufficiently large n and small b, this implies $\|\widehat{\mathbf{H}}_b\mathbf{H}^{-1} - \mathbf{I}_p\|_2 \leq \underline{f}^{-1}r_2 < 1$, and hence

$$\|\mathbf{H}\widehat{\mathbf{H}}_b^{-1} - \mathbf{I}_p\|_2 \leq \frac{r_2}{\underline{f} - r_2}, \quad \|\mathbf{H}\widehat{\mathbf{H}}_b^{-1}\|_2 \leq \frac{\underline{f}}{\underline{f} - r_2}. \tag{B.72}$$

STEP 3 (Controlling the scores). For $\|\mathbf{\Sigma}^{-1/2}\nabla \widehat{Q}_b^G(\boldsymbol{\beta}^*)\|_2$, it follows from Lemma B.2.2 and Proposition B.3.1 that with probability at least $1 - e^{-t}$,

$$\|\mathbf{\Sigma}^{-1/2}\nabla \widehat{Q}_b^G(\boldsymbol{\beta}^*)\|_2 \lesssim \sqrt{\frac{p+t}{n}} + b^{\nu}. \tag{B.73}$$

Turning to $\|\mathbf{\Sigma}^{-1/2}\{\nabla \widehat{Q}_b^G(\overline{\boldsymbol{\beta}}) - \nabla \widehat{Q}_b^G(\boldsymbol{\beta}^*)\} - \mathbf{\Sigma}^{-1/2}\mathbf{J}(\overline{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2$, applying the concentration bound (B.70) and Proposition B.3.3 we obtain that, with probability at least $1 - e^{-t}$ conditioned on $\mathscr{E}_{\text{init}}(t)$,

$$\|\mathbf{\Sigma}^{-1/2}\{\nabla \widehat{Q}_b^G(\overline{\boldsymbol{\beta}}) - \nabla \widehat{Q}_b^G(\boldsymbol{\beta}^*)\} - \mathbf{\Sigma}^{-1/2}\mathbf{J}(\overline{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2 \lesssim \frac{p+t}{nb^{1/2}} + b^{\nu-1}\sqrt{\frac{p+t}{n}} \tag{B.74}$$

as long as $(\frac{p+t}{n})^{1/2} \lesssim b \lesssim 1$.

Finally, combining the bounds (B.70)–(B.74), we conclude that with probability at least

$1-5e^{-t}$,

$$\|\boldsymbol{\Sigma}^{-1/2}\mathbf{J}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)+\boldsymbol{\Sigma}^{-1/2}\nabla\widehat{Q}_b^G(\boldsymbol{\beta}^*)\|_2 \lesssim \left(\sqrt{\frac{p\log n+t}{nb}}+b^{\nu-1}\right)\left(\sqrt{\frac{p+t}{n}}+b^{\nu}\right),$$

provided that $\max\{\frac{p\log n+t}{n}, (\frac{p+t}{n})^{1/2}\} \lesssim b \lesssim 1$. Under the sample size scaling $n \gtrsim p\,(\log n)^2+t$ and bandwidth constraint $b \lesssim (\frac{p+t}{n})^{1/(2\nu)}$, this leads to the claimed bound (B.66). □

Appendix C

Supplementary Material for Chapter 4

C.1 Optimization Algorithms

C.1.1 Low-dimensional setting

Solving the smoothed estimating equations (4.3) and (4.4) is highly similar to Section 3.2. The method is summarized in Algorithm 3.

Algorithm 3. Barzilai-Borwein gradient descent method for minimizing $\widehat{L}_k(\cdot)$.

Input: Censored data $\{(y_i, \boldsymbol{x}_i, \Delta_i)\}_{i=1}^n$, current quantile level $\tau_k \in (0,1)$, previous estimates $\{\widehat{\boldsymbol{\beta}}_j\}_{j=0}^{k-1}$,

initial values $\widehat{\boldsymbol{\beta}}_k^{(0)} = \widehat{\boldsymbol{\beta}}_{k-1}$, bandwidth h, step size upper bound $\alpha_{\max}$, tolerance level ε.

Compute $\widehat{\boldsymbol{\beta}}_k^{(1)} \leftarrow \widehat{\boldsymbol{\beta}}_k^{(0)} - \nabla \widehat{L}_k(\widehat{\boldsymbol{\beta}}_k^{(0)})$

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

  $\boldsymbol{s}_t \leftarrow \widehat{\boldsymbol{\beta}}_k^{(t)} - \widehat{\boldsymbol{\beta}}_k^{(t-1)}$, $\boldsymbol{g}_t \leftarrow \nabla \widehat{L}_k(\widehat{\boldsymbol{\beta}}_k^{(t)}) - \nabla \widehat{L}_k(\widehat{\boldsymbol{\beta}}_k^{(t-1)}) = \nabla \widehat{L}_0(\widehat{\boldsymbol{\beta}}_k^{(t)}) - \nabla \widehat{L}_0(\widehat{\boldsymbol{\beta}}_k^{(t-1)})$

  $\alpha_{t,1} \leftarrow ||\boldsymbol{s}_t||_2^2 / \langle \boldsymbol{s}_t, \boldsymbol{g}_t \rangle$, $\alpha_{t,2} \leftarrow \langle \boldsymbol{s}_t, \boldsymbol{g}_t \rangle / ||\boldsymbol{g}_t||_2^2$

  $\alpha_t \leftarrow \min\{\alpha_{t,1}, \alpha_{t,2}, \alpha_{\max}\}$ if $\langle \boldsymbol{s}_t, \boldsymbol{g}_t \rangle > 0$, and $\alpha_t \leftarrow 1$ otherwise

  Update $\widehat{\boldsymbol{\beta}}_k^{(t+1)} \leftarrow \widehat{\boldsymbol{\beta}}_k^{(t)} - \alpha_t \nabla \widehat{L}_k(\widehat{\boldsymbol{\beta}}_k^{(t)})$

end for when $\|\nabla \widehat{L}_k(\widehat{\boldsymbol{\beta}}_k^{(t)})\|_2 \leq \varepsilon$

C.1.2 High-dimensional setting

In the high dimensional regime, we need to solve the following weighted ℓ_1-penalized programs sequentially:

$$\widehat{\boldsymbol{\beta}}_k = \widehat{\boldsymbol{\beta}}(\tau_k) \in \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \left\{ \widehat{L}_k(\boldsymbol{\beta}) + \|\boldsymbol{\lambda}_k \circ \boldsymbol{\beta}\|_1 \right\}, \;\; k = 0, \ldots, m, \tag{C.1}$$

where $\circ$ denotes the Hadamard product, and $\boldsymbol{\lambda}_k = (\lambda_{k,1}, \ldots, \lambda_{k,p})^{\mathrm{T}}$ may depend on the previous estimates $\{\widehat{\boldsymbol{\beta}}_j\}_{j=0}^{k-1}$. To this end, we apply the iterative local adaptive majorize-minimize (I-LAMM) algorithm proposed in Fan et al. [2018].

To illustrate the basic ideas, consider the general problem of minimizing a nonlinear function $f(\cdot)$ on $\mathbb{R}^p$. Starting at a given point $\boldsymbol{\beta}_0$, the majorize-minimize (MM) algorithm involves two steps: first, construct a majorizing function $g(\cdot \,|\, \boldsymbol{\beta}_0)$, satisfying

$$g(\boldsymbol{\beta}_0 \,|\, \boldsymbol{\beta}_0) = f(\boldsymbol{\beta}_0) \;\text{ and }\; \underbrace{g(\boldsymbol{\beta} \,|\, \boldsymbol{\beta}_0) \geq f(\boldsymbol{\beta}) \;\text{ for any } \boldsymbol{\beta} \in \mathbb{R}^p}_{\text{global majorization property}};$$

secondly, update $\boldsymbol{\beta}_0$ by $\boldsymbol{\beta}_1 := \operatorname{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} g(\boldsymbol{\beta} \,|\, \boldsymbol{\beta}_0)$ [Lange, Hunter and Yang, 2000]. Noting that

$$f(\boldsymbol{\beta}_1) \overset{\text{majorization}}{\leq} g(\boldsymbol{\beta}_1 \,|\, \boldsymbol{\beta}_0) \overset{\text{minimization}}{\leq} g(\boldsymbol{\beta}_0 \,|\, \boldsymbol{\beta}_0) = f(\boldsymbol{\beta}_0),$$

the objective value of such an algorithm is non-increasing in each step. In fact, the global majorization property is not necessary to ensure non-increasing objective values. Instead, we only need the following local majorization property

$$g(\boldsymbol{\beta}_0 \,|\, \boldsymbol{\beta}_0) = f(\boldsymbol{\beta}_0) \;\text{ and }\; g(\boldsymbol{\beta}_1 \,|\, \boldsymbol{\beta}_0) \geq f(\boldsymbol{\beta}_1).$$

To construct a proper majorizing function for $\widehat{L}_k(\cdot)$ around $\boldsymbol{\beta}_0$, we define an isotropic

quadratic function

$$F(\boldsymbol{\beta};\phi,\boldsymbol{\beta}_0) := \widehat{L}_k(\boldsymbol{\beta}_0) + \langle \nabla \widehat{L}_k(\boldsymbol{\beta}_0), \boldsymbol{\beta}-\boldsymbol{\beta}_0\rangle + \frac{\phi}{2}\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|_2^2$$

for some $\phi > 0$. It is easy to see that $F(\boldsymbol{\beta}_0;\phi,\boldsymbol{\beta}_0) = \widehat{L}_k(\boldsymbol{\beta}_0)$. Using such a surrogate loss function, the weighted ℓ_1-penalized program $\min_{\boldsymbol{\beta}\in\mathbb{R}^p}\{F(\boldsymbol{\beta};\phi,\boldsymbol{\beta}_0)+\|\boldsymbol{\lambda}_k \circ \boldsymbol{\beta}\|_1\}$ admits a closed-form solution $\boldsymbol{\beta}_1 = S_{\text{soft}}(\boldsymbol{\beta}_0 - \nabla\widehat{L}_k(\boldsymbol{\beta}_0)/\phi, \boldsymbol{\lambda}_k/\phi)$, where $S_{\text{soft}}(\boldsymbol{\beta},\boldsymbol{\lambda}) := (\text{sign}(\beta_j)\max\{|\beta_j| - \lambda_{k,j}, 0\})_{j=1,\ldots,p}$ is the soft-thresholding operator. Moreover, the quadratic coefficient $\phi > 0$ should be sufficiently large so that the local majorization property $F(\boldsymbol{\beta}_1;\phi,\boldsymbol{\beta}_0) \geq \widehat{L}_k(\boldsymbol{\beta}_1)$ is satisfied. Starting with a relatively small value $\phi = \phi_0$, we iteratively increase ϕ by a factor of $\gamma > 1$ and compute

$$\boldsymbol{\beta}_{1,\ell} = S_{\text{soft}}(\boldsymbol{\beta}_0 - \nabla\widehat{L}_k(\boldsymbol{\beta}_0)/\phi_\ell, \boldsymbol{\lambda}_k/\phi_\ell) \;\text{ with }\; \phi_\ell = \gamma^\ell\phi_0,\;\; \ell = 0,1,\ldots.$$

until the local majorization property holds. Repeating this procedure yields a sequence of iterates $\{\boldsymbol{\beta}_t\}_{t=0,1,\ldots}$ until the stopping criterion is met, say $\|\boldsymbol{\beta}_{t+1}-\boldsymbol{\beta}_t\|_2 \leq \varepsilon$. We treat $\phi_0,\gamma,\varepsilon$ as internal optimization parameters; the default choice is $(\phi_0,\gamma,\varepsilon) = (0.5, 1.5, 10^{-5})$.

C.2 Proofs of the Main Results in Section 4.3.2

To begin with, we revisit and define some notations that will be frequently used. For predetermined grid points $\tau_L = \tau_0 < \tau_1 < \cdots < \tau_m = \tau_U$, we write $\boldsymbol{\beta}_j^* = \boldsymbol{\beta}^*(\tau_j)$ and $\widehat{\boldsymbol{\beta}}_j = \widehat{\boldsymbol{\beta}}(\tau_j)$, $j = 0,\ldots,m$. Since the estimators $\{\widehat{\boldsymbol{\beta}}_j\}_{j=0}^m$ are constructed sequentially, the statistical error of $\widehat{\boldsymbol{\beta}}_j$ at quantile level τ_j depends on the accumulated errors of $\widehat{\boldsymbol{\beta}}_0,\ldots,\widehat{\boldsymbol{\beta}}_{j-1}$.

For every $r > 0$, define the local ellipse $\Theta(r) = \{\boldsymbol{\delta}\in\mathbb{R}^p : \|\boldsymbol{\delta}\|_\Sigma \leq r\}$ under the Σ-induced norm. Under Condition 4.3.2 on the (random) feature vector $\boldsymbol{x}\in\mathbb{R}^p$, for every $\delta\in(0,1]$ we set

$$\eta_\delta = \inf\big\{\eta > 0 : \mathbb{E}\big\{(\boldsymbol{z}^{\mathrm{T}}\boldsymbol{v})^2\mathbb{1}(|\boldsymbol{z}^{\mathrm{T}}\boldsymbol{v}| > \eta)\big\} \leq \delta \;\text{ for all }\boldsymbol{v}\in\mathbb{S}^{p-1}\big\}, \tag{C.2}$$

where $\boldsymbol{z} = \Sigma^{-1/2}\boldsymbol{x}$. Since $\mathbb{E}(\boldsymbol{z}^{\mathrm{T}}\boldsymbol{v})^2 = 1$ for any $\boldsymbol{v} \in \mathbb{S}^{p-1}$, η_δ is well-defined for each δ, and depends implicitly on the underlying distribution of $\boldsymbol{z}$. Throughout the proof, we write

$$\boldsymbol{z}_i = (z_{i1}, \ldots, z_{ip})^{\mathrm{T}} = \Sigma^{-1/2}\boldsymbol{x}_i,\ i = 1, \ldots, n.$$

For a non-negative kernel function $K(\cdot)$ and a bandwidth $h > 0$, we write

$$K_h(u) = h^{-1}K(u/h),\ \ \bar{K}_h(u) = \bar{K}(u/h) \quad \text{and} \quad \bar{K}(u) = \int_{-\infty}^{u} K(t)\,\mathrm{d}t.$$

C.2.1 Technical lemmas

We first collect several technical lemmas that serve as the building blocks for proving the main results.

Lemma C.2.1 (Convexity lemma)**.** *For a vector-valued function $Q(\boldsymbol{\beta}) : \mathbb{R}^p \to \mathbb{R}^p$ with positive semi-definite Jacobian, define the corresponding divergence function $D = D_Q : \mathbb{R}^p \times \mathbb{R}^p \to [0, \infty)$ as $D(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = \langle Q(\boldsymbol{\beta}_1) - Q(\boldsymbol{\beta}_2), \boldsymbol{\beta}_1 - \boldsymbol{\beta}_2 \rangle$. For any $\boldsymbol{\beta}, \boldsymbol{\beta}' \in \mathbb{R}^p$ and $\eta \in [0,1]$, we have*

$$D(\boldsymbol{\beta}' + \eta(\boldsymbol{\beta} - \boldsymbol{\beta}'), \boldsymbol{\beta}') \leq \eta D(\boldsymbol{\beta}, \boldsymbol{\beta}').$$

Lemma C.2.2 below provides a Bernstein-type inequality for the ℓ_2-norm of a sum of centered random vectors, which will be frequently used to bound the smoothed estimating functions.

Lemma C.2.2. *Assume Condition 4.3.2 holds, and let $\{\xi_i\}_{i=1}^n$ be independent random variables satisfying $\mathbb{E}(\xi_i^2 | \boldsymbol{x}_i) \leq \sigma^2$ and $|\xi_i| \leq M$ for some $M \geq \sigma > 0$. Then, for any $t > 0$,*

$$\left\| \frac{1}{n} \sum_{i=1}^{n} (\xi_i \boldsymbol{z}_i - \mathbb{E}\,\xi_i \boldsymbol{z}_i) \right\|_2 \leq 2\sigma \sqrt{\frac{p}{n}} + \sigma \sqrt{\frac{2t}{n}} + M\zeta_p \frac{4t}{3n}$$

holds with probability at least $1 - e^{-t}$, where $\zeta_p = \max_{\boldsymbol{x} \in \mathscr{X}} \|\boldsymbol{x}\|_{\Sigma^{-1}}$.

Lemma C.2.3 provides concentration inequalities for some of the stochastic terms in the estimating functions $\widehat{Q}_0(\cdot)$ and $\widehat{Q}_j(\cdot)$ ($j \geq 1$).

Lemma C.2.3. *Let $j = 0, 1, \ldots, m$ and $t > 0$.*

(i) With probability at least $1 - e^{-t}$,

$$||\widehat{Q}_0(\boldsymbol{\beta}_j^*) - \mathbb{E}\widehat{Q}_0(\boldsymbol{\beta}_j^*)||_{\Sigma^{-1}} \leq 2\bar{\tau}_0\left(\sqrt{\frac{p}{n}} + +\sqrt{\frac{t}{2n}} + \zeta_p\frac{2t}{3n}\right),$$

where $\bar{\tau}_0 = \max(\tau_0, 1-\tau_0)$.

(ii) With probability at least $1 - e^{-t}$,

$$\left\|\frac{1}{n}\sum_{i=1}^{n}(1-\mathbb{E})\bar{K}_h(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_j^*)\boldsymbol{x}_i\right\|_{\Sigma^{-1}} \leq 2\left(\sqrt{\frac{p}{n}} + \sqrt{\frac{t}{2n}} + \zeta_p\frac{2t}{3n}\right).$$

(iii) With probability at least $1 - e^{-t}$,

$$\begin{aligned}&\left\|\frac{1}{n}\sum_{i=1}^{n}(1-\mathbb{E})\int_{\tau_j}^{\tau_{j+1}}\left\{\bar{K}_h(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*(u)) - \bar{K}_h(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*(\tau_j))\right\}\mathrm{d}H(u)\cdot\boldsymbol{x}_i\right\|_{\Sigma^{-1}}\\ &\lesssim m_4^{1/2}w_j\cdot\delta^*\sqrt{\frac{p+t}{nh}} + w_j\cdot\zeta_p^2\delta^*\frac{t}{nh},\end{aligned}$$

where $w_j = H(\tau_{j+1}) - H(\tau_j) = \log(\frac{1-\tau_j}{1-\tau_{j+1}})$.

The following lemma concerns the first-order property of the smoothed estimating functions $\widehat{Q}_j(\cdot)$ in (4.3) and (4.4). Define the corresponding symmetrized Bregman divergence $D : \mathbb{R}^p \times \mathbb{R}^p \to [0, \infty)$ as

$$D(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) := \langle\widehat{Q}_j(\boldsymbol{\beta}_1) - \widehat{Q}_j(\boldsymbol{\beta}_2), \boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\rangle = \langle\widehat{Q}_0(\boldsymbol{\beta}_1) - \widehat{Q}_0(\boldsymbol{\beta}_2), \boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\rangle. \tag{C.3}$$

Note that the divergence D is independent of j.

Lemma C.2.4 (Restricted strong convexity)**.** *Assume Conditions 3.1–3.3 hold, and let $h, r > 0$ satisfy $4\eta_{1/4} r \le h \le 1$ with $\eta_{1/4}$ defined in* (C.2)*. Then, for any $0 \le j \le m$ and $t > 0$,*

$$\inf_{\boldsymbol{\beta} \in \boldsymbol{\beta}_j^* + \Theta(r)} \frac{D(\boldsymbol{\beta}, \boldsymbol{\beta}_j^*)}{\kappa_l \, ||\boldsymbol{\beta} - \boldsymbol{\beta}_j^*||_\Sigma^2} \ge \frac{3}{4}\underline{g} - \overline{g}^{1/2} r^{-1} \left(\frac{5}{4}\sqrt{\frac{hp}{n}} + \sqrt{\frac{ht}{8n}} \right) - c r^{-2} \frac{ht}{n}$$

with probability at least $1 - e^{-t}$, where $c = 1/4 + 1/48 \approx 0.27$.

For $j = 0, 1, \ldots$ and $r > 0$, define

$$\varpi_j(r) = \sup_{\boldsymbol{\beta} \in \boldsymbol{\beta}_j^* + \Theta(r)} \left\| \frac{1}{n} \sum_{i=1}^n \big\{ \bar{K}_h(y_i - \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}) - \bar{K}_h(y_i - \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_j^*) \big\} \boldsymbol{x}_i + \mathbf{H}_j (\boldsymbol{\beta} - \boldsymbol{\beta}_j^*) \right\|_{\Sigma^{-1}}, \tag{C.4}$$

$$\omega_j(r) = \sup_{\boldsymbol{\beta} \in \boldsymbol{\beta}_j^* + \Theta(r)} \left\| \frac{1}{n} \sum_{i=1}^n \Delta_i \big\{ \bar{K}_h(\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta} - y_i) - \bar{K}_h(\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_j^* - y_i) \big\} \boldsymbol{x}_i - \mathbf{J}_j (\boldsymbol{\beta} - \boldsymbol{\beta}_j^*) \right\|_{\Sigma^{-1}}, \tag{C.5}$$

where $\mathbf{J}_j = \mathbf{J}(\tau_j) = \mathbb{E}\{ g(\boldsymbol{x}^{\mathrm{T}} \boldsymbol{\beta}_j^* | \boldsymbol{x}) \boldsymbol{x} \boldsymbol{x}^{\mathrm{T}} \}$ and $\mathbf{H}_j = \mathbf{H}(\tau_j) = \mathbb{E}\{ f(\boldsymbol{x}^{\mathrm{T}} \boldsymbol{\beta}_j^* | \boldsymbol{x}) \boldsymbol{x} \boldsymbol{x}^{\mathrm{T}} \}$. The following lemma provides upper bounds for the two suprema $\varpi_j(r)$ and $\omega_j(r)$ for any given $r > 0$.

Lemma C.2.5. *Assume Conditions 4.3.1–4.3.3 hold, and write $m_k = \sup_{\boldsymbol{u} \in \mathbb{S}^{p-1}} \mathbb{E}|\boldsymbol{z}^{\mathrm{T}} \boldsymbol{u}|^k$ ($k = 3, 4$). Let $j = 0, 1, \ldots, m$ and $r > 0$.*

(i) With probability at least $1 - e^{-t}$,

$$\begin{aligned} &\sup_{\boldsymbol{\beta} \in \boldsymbol{\beta}_j^* + \Theta(r)} \left\| \frac{1}{n} \sum_{i=1}^n (1 - \mathbb{E}) \big\{ \bar{K}_h(y_i - \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}) - \bar{K}_h(y_i - \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_j^*) \big\} \boldsymbol{x}_i \right\|_{\Sigma^{-1}} \\ &\qquad \lesssim (\kappa_u \overline{f} m_4)^{1/2} \sqrt{\frac{p+t}{nh}} \cdot r, \end{aligned} \tag{C.6}$$

provided that the "effective" sample size satisfies $nh \gtrsim \zeta_p^2 (p + t)$. Moreover, for any $\boldsymbol{\beta} \in \boldsymbol{\beta}_j^* + \Theta(r)$,

$$\left\| \mathbb{E}\big\{ \bar{K}_h(y - \boldsymbol{x}^{\mathrm{T}} \boldsymbol{\beta}) - \bar{K}_h(y - \boldsymbol{x}^{\mathrm{T}} \boldsymbol{\beta}_j^*) \big\} \boldsymbol{x} + \mathbf{H}_j (\boldsymbol{\beta} - \boldsymbol{\beta}_j^*) \right\|_{\Sigma^{-1}} \le l_1 \big(0.5 m_3 r + \kappa_1 h \big) \cdot r.$$

(ii) With probability at least $1-e^{-t}$,

$$\sup_{\boldsymbol{\beta}\in\boldsymbol{\beta}_j^*+\Theta(r)} \left\| \frac{1}{n}\sum_{i=1}^{n}(1-\mathbb{E})\Delta_i\big\{\bar{K}_h(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}-y_i)-\bar{K}_h(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_j^*-y_i)\big\}\boldsymbol{x}_i \right\|_{\Sigma^{-1}} \lesssim (\kappa_u \overline{g} m_4)^{1/2}\sqrt{\frac{p+t}{nh}}\cdot r, \tag{C.7}$$

provided that $nh \gtrsim \zeta_p^2(p+t)$. *Moreover, for any* $\boldsymbol{\beta}\in\boldsymbol{\beta}_j^*+\Theta(r)$,

$$\left\| \mathbb{E}\Delta_i\big\{\bar{K}_h(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}-y)-\bar{K}_h(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_j^*-y)\big\}\boldsymbol{x}-\mathbf{J}_j(\boldsymbol{\beta}-\boldsymbol{\beta}_j^*) \right\|_{\Sigma^{-1}} \le l_1\big(0.5m_3 r+\kappa_1 h\big)\cdot r.$$

Lemma C.2.6. *Assume Conditions 4.3.1–4.3.3 hold. For any* $\tau_L \le \tau_l < \tau_u \le \tau_U$,

$$\sup_{\tau\in[\tau_l,\tau_u]} \left\| \frac{1}{n}\sum_{i=1}^{n}(1-\mathbb{E})\int_{\tau_l}^{\tau}\bar{K}_h(y_i-\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*(u))\mathrm{d}H(u)\cdot\boldsymbol{x}_i \right\|_{\Sigma^{-1}} \lesssim \frac{\tau_u-\tau_l}{1-\tau_u}\left(\sqrt{\frac{p+\log n+t}{n}}+\zeta_p\frac{\log n+t}{n}\right)$$

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

Moreover, Lemma C.2.7 provides upper bounds on the approximation error, which consists of the smoothing and discretization errors. Let $Q_0(\boldsymbol{\beta})=\mathbb{E}\widehat{Q}_0(\boldsymbol{\beta})$.

Lemma C.2.7. *For each* $j=1,\ldots,m$,

$$\left\| Q_0(\boldsymbol{\beta}_j^*)-\sum_{\ell=0}^{j-1} w_\ell \mathbb{E}\{\bar{K}_h(y-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_\ell^*)\boldsymbol{x}\} \right\|_{\Sigma^{-1}} \le \frac{1}{2}l_1\kappa_2 h^2\big\{1+H(\tau_j)-H(\tau_0)\big\}+(\overline{f}/\underline{f})\sum_{\ell=0}^{j-1} w_\ell(\tau_{\ell+1}-\tau_\ell), \tag{C.8}$$

where $w_\ell = H(\tau_{\ell+1})-H(\tau_\ell)$. *In particular,* $\|Q_0(\boldsymbol{\beta}_0^*)\|_{\Sigma^{-1}} \le 0.5 l_1\kappa_2 h^2$.

The next lemma establishes the asymptotic uniform equicontinuity of the centered process $\mathbb{G}_n(\cdot)$ in $\ell^\infty([\tau_L,\tau_U])$. This is an equivalent definition to asymptotic tightness, and is an important

step towards the weak convergence stated in Theorem 4.3.3.

Lemma C.2.8 (Asymptotic uniform equicontinuity)**.** *Assume the conditions of Theorem 4.3.3 hold, and let* $\{\boldsymbol{a}_n\}_{n\geq 1}$ *be a normalized sequence such that* $\|\boldsymbol{a}_n\|_\Sigma = 1$*. Then, for any* $x > 0$*,*

$$\lim_{\delta\to 0}\limsup_{n\to\infty}\mathbb{P}\left\{\sup_{|\tau_1-\tau_2|<\delta}|\mathbb{G}_n(\tau_1)-\mathbb{G}_n(\tau_2)|>x\right\}=0,$$

where $\mathbb{G}_n(\cdot)$ *is defined in* (4.20)*.*

C.2.2 Proof of Theorem 4.3.1

We first prove a uniform upper bound over the grid of τ-values—$\{\tau_0,\tau_1,\ldots,\tau_m\}$. That is, with probability at least $1-C_1n^{-1}$,

$$\|\widehat{\boldsymbol{\beta}}_j-\boldsymbol{\beta}_j^*\|_\Sigma=\|\widehat{\boldsymbol{\beta}}(\tau_j)-\boldsymbol{\beta}^*(\tau_j)\|_\Sigma\leq r_j,\quad j=0,1,\ldots,m \tag{C.9}$$

for a sequence of radii $\{r_j\}_{j=0,1,\ldots,m}$ and some absolute constant $C_1>0$. To begin with, define a crude (sub-optimal) convergence radius $r^\diamondsuit = h/(4\eta_{1/4})$ with $\eta_{1/4}$ given in (C.2). Accordingly, define "intermediate" points $\widetilde{\boldsymbol{\beta}}_j=(1-u_j)\boldsymbol{\beta}_j^*+u_j\widehat{\boldsymbol{\beta}}_j$, where

$$u_j=\sup\left\{u\in[0,1]:u(\widehat{\boldsymbol{\beta}}_j-\boldsymbol{\beta}_j^*)\in\Theta(r^\diamondsuit)\right\}\begin{cases}=1 & \text{if } \widehat{\boldsymbol{\beta}}_j\in\boldsymbol{\beta}_j^*+\Theta(r^\diamondsuit),\\ \in(0,1) & \text{if } \widehat{\boldsymbol{\beta}}_j\notin\boldsymbol{\beta}_j^*+\Theta(r^\diamondsuit).\end{cases}$$

It is easy to see that $\widetilde{\boldsymbol{\beta}}_j=\widehat{\boldsymbol{\beta}}_j$ if $\widehat{\boldsymbol{\beta}}_j\in\boldsymbol{\beta}_j^*+\Theta(r^\diamondsuit)$, and $\widetilde{\boldsymbol{\beta}}_j\in\boldsymbol{\beta}_j^*+\partial\Theta(r^\diamondsuit)$ if $\widehat{\boldsymbol{\beta}}_j\notin\boldsymbol{\beta}_j^*+\Theta(r_j)$. Here $\partial\Theta(r)$ denotes the boundary of $\Theta(r)$. In either case, we have $\widetilde{\boldsymbol{\beta}}_j\in\boldsymbol{\beta}_j^*+\Theta(r^\diamondsuit)$.

Recall the symmetrized Bregman divergence $D(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2)=\langle\widehat{Q}_j(\boldsymbol{\beta}_1)-\widehat{Q}_j(\boldsymbol{\beta}_2),\boldsymbol{\beta}_1-\boldsymbol{\beta}_2\rangle$ defined in (C.3). Applying Lemma C.2.1 yields that, for each $j=0,1,\ldots,m$,

$$D(\widetilde{\boldsymbol{\beta}}_j,\boldsymbol{\beta}_j^*)\leq u_j\cdot D(\widehat{\boldsymbol{\beta}}_j,\boldsymbol{\beta}_j^*)=u_j\cdot\langle\widehat{Q}_j(\widehat{\boldsymbol{\beta}}_j)-\widehat{Q}_j(\boldsymbol{\beta}_j^*),\widehat{\boldsymbol{\beta}}_j-\boldsymbol{\beta}_j^*\rangle.$$

Since $\widehat{\boldsymbol{\beta}}_j$ solves the estimating equation $\widehat{Q}_j(\widehat{\boldsymbol{\beta}}_j) = \mathbf{0}$, by the Cauchy–Schwartz inequality we have

$$D(\widetilde{\boldsymbol{\beta}}_j, \boldsymbol{\beta}_j^*) \le u_j \cdot \langle -\widehat{Q}_j(\boldsymbol{\beta}_j^*), \widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^* \rangle \le \|\widehat{Q}_j(\boldsymbol{\beta}_j^*)\|_{\Sigma^{-1}} \cdot \|\widetilde{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^*\|_{\Sigma}.$$

For some curvature parameter $\kappa > 0$ to be specified, define the event

$$\mathscr{F} = \bigcap_{j=0}^{m} \left\{ D(\boldsymbol{\beta}, \boldsymbol{\beta}_j^*) \ge \kappa \cdot \|\boldsymbol{\beta} - \boldsymbol{\beta}_j^*\|_{\Sigma}^2 \text{ for all } \boldsymbol{\beta} \in \boldsymbol{\beta}_j^* + \Theta(r^{\diamondsuit}) \right\}. \tag{C.10}$$

Conditioning on $\mathscr{F}$, it follows that

$$\|\widetilde{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^*\|_{\Sigma} \le \kappa^{-1} \|\widehat{Q}_j(\boldsymbol{\beta}_j^*)\|_{\Sigma^{-1}}, \quad j = 0, 1, \ldots, m. \tag{C.11}$$

Next we derive upper bounds for $\{\|\widehat{Q}_j(\boldsymbol{\beta}_j^*)\|_{\Sigma^{-1}}\}_{j=0,1,\ldots,m}$ sequentially. For each j, we decompose $\widehat{Q}_j(\boldsymbol{\beta}_j^*)$ as

$$\widehat{Q}_j(\boldsymbol{\beta}_j^*) = \widehat{Q}_0(\boldsymbol{\beta}_j^*) - Q_0(\boldsymbol{\beta}_j^*) - \sum_{\ell=0}^{j-1} w_\ell (\widehat{\Delta}_\ell + \Delta_\ell) + Q_0(\boldsymbol{\beta}_j^*) - \sum_{\ell=0}^{j-1} w_\ell \mathbb{E}\{\bar{K}_h(y - \boldsymbol{x}^{\mathrm{T}} \boldsymbol{\beta}_\ell^*)\boldsymbol{x}\}$$

where $Q_0(\boldsymbol{\beta}) = \mathbb{E}\widehat{Q}_0(\boldsymbol{\beta})$, $w_\ell = H(\tau_{\ell+1}) - H(\tau_\ell)$,

$$\widehat{\Delta}_\ell = \frac{1}{n} \sum_{i=1}^{n} \{\bar{K}_h(y_i - \boldsymbol{x}_i^{\mathrm{T}} \widehat{\boldsymbol{\beta}}_\ell) - \bar{K}_h(y_i - \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_\ell^*)\} \boldsymbol{x}_i \text{ and } \Delta_\ell = \frac{1}{n} \sum_{i=1}^{n} (1 - \mathbb{E}) \bar{K}_h(y_i - \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_\ell^*) \boldsymbol{x}_i. \tag{C.12}$$

By the triangle inequality,

$$\|\widehat{Q}_j(\boldsymbol{\beta}_j^*)\|_{\Sigma^{-1}} \le \underbrace{\|\widehat{Q}_0(\boldsymbol{\beta}_j^*) - Q_0(\boldsymbol{\beta}_j^*)\|_{\Sigma^{-1}}}_{\text{statistical error of the } j\text{th estimating equation}} + \underbrace{\sum_{\ell=0}^{j-1} w_\ell \big(\|\widehat{\Delta}_\ell\|_{\Sigma^{-1}} + \|\Delta_\ell\|_{\Sigma^{-1}}\big)}_{\text{accumulated error}} \tag{C.13}$$
$$+ \underbrace{\Big\| Q_0(\boldsymbol{\beta}_j^*) - \sum_{\ell=0}^{j-1} w_\ell \mathbb{E}\{\bar{K}_h(y - \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_\ell^*)\boldsymbol{x}\} \Big\|_{\Sigma^{-1}}}_{\text{approximation error}}, \quad j = 1, \ldots, m.$$

In particular, $\|\widehat{Q}_0(\boldsymbol{\beta}_0^*)\|_{\Sigma^{-1}} \le \|\widehat{Q}_0(\boldsymbol{\beta}_0^*) - Q_0(\boldsymbol{\beta}_0^*)\|_{\Sigma^{-1}} + \|Q_0(\boldsymbol{\beta}_0^*)\|_{\Sigma^{-1}}$. For the approximation error term in (C.13), by Lemma C.2.7 we have

$$\begin{aligned} &\Big\| Q_0(\boldsymbol{\beta}_j^*) - \sum_{\ell=0}^{j-1} w_\ell \mathbb{E}\{\bar{K}_h(y - \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_\ell^*)\boldsymbol{x}\} \Big\|_{\Sigma^{-1}} \\ &\quad \le \frac{1}{2} l_1 \kappa_2 h^2 + \sum_{\ell=0}^{j-1} w_j a < a + \sum_{\ell=0}^{j-1} w_j a \quad \text{with} \quad a := \frac{1}{2} l_1 \kappa_2 h^2 + \overline{f} \underline{f}^{-1} \delta^*. \end{aligned} \tag{C.14}$$

For some $\delta > 0$ to be determined, define the second event

$$\mathscr{G} = \left\{ \max_{0 \le j \le m} \|\widehat{Q}_0(\boldsymbol{\beta}_j^*) - Q_0(\boldsymbol{\beta}_j^*)\|_{\Sigma^{-1}} \bigvee \max_{0 \le \ell \le m-1} \|\Delta_\ell\|_{\Sigma^{-1}} \le \delta \right\}, \tag{C.15}$$

where the Δ_ℓ's are given in (C.12). Conditioned on $\mathscr{F} \cap \mathscr{G}$, it follows from (C.13)–(C.15) that

$$\|\widehat{Q}_0(\boldsymbol{\beta}_0^*)\|_{\Sigma^{-1}} < \delta + a, \quad \|\widehat{Q}_j(\boldsymbol{\beta}_j^*)\|_{\Sigma^{-1}} < \delta + a + \sum_{\ell=0}^{j-1} w_\ell \big(\delta + a + \|\widehat{\Delta}_\ell\|_{\Sigma^{-1}}\big), \quad j = 1, \ldots, m.$$

Based on the above general bounds, we iteratively control $\|\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^*\|_\Sigma$ and $\|\widehat{\Delta}_j\|_{\Sigma^{-1}}$, starting at $j = 0$. By (C.11),

$$\|\widetilde{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}_0^*\|_\Sigma \le \kappa^{-1} \|\widehat{Q}_0(\boldsymbol{\beta}_0^*)\|_{\Sigma^{-1}} < r_0 := \kappa^{-1}(\delta + a).$$

Provided $r_0 \le r^\diamondsuit$, the intermediate point $\widetilde{\boldsymbol{\beta}}_0$ falls into the interior of the local region $\boldsymbol{\beta}_0^* + \Theta(r^\diamondsuit)$.

Via proof by contradiction, we must have $\widehat{\boldsymbol{\beta}}_0 = \widetilde{\boldsymbol{\beta}}_0$ and hence $\widehat{\boldsymbol{\beta}}_0 \in \boldsymbol{\beta}_0^* + \Theta(r_0)$.

Turning to $(\widetilde{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_1)$, it follows from (C.11) with $j = 1$ that $\|\widetilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*\|_\Sigma < \kappa^{-1}\{\delta + a + w_0(\|\widehat{\Delta}_0\|_{\Sigma^{-1}} + \delta + a)\}$. Note that $\widehat{\Delta}_0$ depends on the preceding estimate $\widehat{\boldsymbol{\beta}}_0$. Since, as proved in the last step, $\widehat{\boldsymbol{\beta}}_0 \in \boldsymbol{\beta}_0^* + \Theta(r_0)$ conditioned on $\mathscr{F} \cap \mathscr{G}$, it follows that

$$\|\widehat{\Delta}_0\|_{\Sigma^{-1}} \le \varpi_0(r_0) + \|\mathbf{H}(\tau_0)(\widehat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}_0^*)\|_{\Sigma^{-1}} \le \varpi_0(r_0) + \overline{f} r_0,$$

where $\varpi_0(\cdot)$ is defined in (C.4). Conditioned further on $\{\varpi_0(r_0) \le \overline{f} r_0\}$, this implies

$$\|\widetilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*\|_\Sigma < r_1 := \kappa^{-1}\big\{\delta + a + w_0(2\overline{f} r_0 + \delta + a)\big\}.$$

As long as $r_1 \le r^\diamondsuit$, $\widetilde{\boldsymbol{\beta}}_1$ lies in the interior of $\boldsymbol{\beta}_1^* + \Theta(r^\diamondsuit)$, which enforces $\widehat{\boldsymbol{\beta}}_1 = \widetilde{\boldsymbol{\beta}}_1$ and hence $\widehat{\boldsymbol{\beta}}_1 \in \boldsymbol{\beta}_1^* + \Theta(r_1)$.

Applying the above argument repeatedly, at the j-th step $(1 \le j \le m)$, we obtain that conditioned on $\{\varpi_{j-1}(r_{j-1}) \le \overline{f} r_{j-1}\}$,

$$\begin{aligned} \|\widetilde{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^*\|_\Sigma &< \frac{1}{\kappa}\left\{\delta + a + \sum_{\ell=0}^{j-1} w_j\big(\|\widehat{\Delta}_\ell\|_{\Sigma^{-1}} + \delta + a\big)\right\} \\ &\le \frac{1}{\kappa}\left\{\delta + a + \sum_{\ell=0}^{j-1} w_\ell(2\overline{f} r_\ell + \delta + a)\right\} =: r_j. \end{aligned} \tag{C.16}$$

Provided $r_j \le r^\diamondsuit$, by way of contradiction we must have $\widehat{\boldsymbol{\beta}}_j = \widetilde{\boldsymbol{\beta}}_j \in \boldsymbol{\beta}_j^* + \Theta(r_j)$. Equivalently, the above sequence of radii $\{r_j\}_{j=0}^m$ can be recursively defined as

$$r_j = \big(1 + 2\kappa^{-1}\overline{f} w_{j-1}\big) r_{j-1} + \kappa^{-1} w_{j-1}(\delta + a), \;\; j = 1, \ldots, m, \;\text{ and }\; r_0 = \kappa^{-1}(\delta + a),$$

where a is given in (C.14). Taking $C = \kappa^{-1}(2\overline{f}+1)$, it follows that

$$r_j \le (1+Cw_{j-1})r_{j-1} \le \cdots \le \prod_{\ell=0}^{j-1}(1+Cw_\ell)\cdot r_0 \le \exp\left(C\sum_{\ell=0}^{j-1} w_\ell\right)\cdot r_0 = \left(\frac{1-\tau_L}{1-\tau_j}\right)^C \cdot r_0. \tag{C.17}$$

Thus far we have established the result $\widehat{\boldsymbol{\beta}}_j \in \boldsymbol{\beta}_j^* + \Theta(r_j)$ $(j = 0,1,\ldots,m)$ as a deterministic claim, but conditioned on the "good" event

$$\mathscr{F} \cap \mathscr{G} \cap \bigcap_{\ell=0}^{m-1} \{\varpi_\ell(r_\ell) \le \overline{f} r_\ell\}$$

with properly chosen κ, δ and $\{r_j\}_{j=0}^m$. By Lemmas C.2.4 and C.2.3, we choose $\kappa = (\underline{g}\kappa_l)/2$ and $\delta \asymp \sqrt{(p+\log n)/n} + \zeta_p \log(n)/n$, so that $\mathbb{P}(\mathscr{F}^c) \le (m+1)/n^2$ and $\mathbb{P}(\mathscr{G}^c) \le 2(m+1)/n^2$ as long as $nh \gtrsim p + \log n$. With this choice of δ, and since $a = 0.5 l_1 \kappa_2 h^2 + (\overline{f}/\underline{f})\delta^* \lesssim h^2 + n^{-1/2}$, we obtain from (C.17) that

$$r_j \le \left(\frac{1-\tau_L}{1-\tau_j}\right)^C \cdot r_0 \asymp \left(\frac{1-\tau_L}{1-\tau_j}\right)^C \underline{g}^{-1}\left(\sqrt{\frac{p+\log n}{n}} + \zeta_p \frac{\log n}{n} + h^2\right).$$

Moreover, it follows from Lemma C.2.5 that with probability at least $1 - m/n^2$,

$$\varpi_\ell(r_\ell) \lesssim \left(m_4^{1/2}\sqrt{\frac{p+\log n}{nh}} + m_3 r_\ell + h\right)\cdot r_\ell \ \text{ for all } \ \ell = 0,1,\ldots,m-1,$$

provided $nh \gtrsim \zeta_p^2 (p+\log n)^{1/2}$. By the prescribed choice of the bandwidth $h = h_n \asymp \{(p+\log n)/n\}^\gamma$ with $\gamma \in [1/4, 1/2)$, and the requirement $n \gtrsim \zeta_p^{2/(1-\gamma)}(p+\log n)^{(1/2-\gamma)/(1-\gamma)}$, we conclude that the above "good" event occurs with probability at least $1 - C_1 n^{-1}$, and

$$\left(\frac{1-\tau_L}{1-\tau_j}\right)^C \underline{g}^{-1}\sqrt{\frac{p+\log n}{n}} \asymp r_j \le r^\diamondsuit \asymp \left(\frac{p+\log n}{n}\right)^\gamma \ \text{ for all } j = 0,1,\ldots,m.$$

This proves the claim (C.9).

To establish the uniform rate of convergence for $\{\widehat{\boldsymbol{\beta}}(\tau), \tau \in [\tau_L, \tau_U]\}$, define disjoint intervals $\mathscr{I}_j = [\tau_j, \tau_{j+1})$ for $j = 0, 1, \ldots, m-1$, and $\mathscr{I}_m = \{\tau_m\}$. For any $\tau \in [\tau_L, \tau_U]$, by the definition of $\widehat{\boldsymbol{\beta}}(\cdot)$ there exits a unique index $j \in \{0, 1, \ldots, m\}$ such that $\tau \in \mathscr{I}_j$ and $\widehat{\boldsymbol{\beta}}(\tau) = \widehat{\boldsymbol{\beta}}(\tau_j) = \widehat{\boldsymbol{\beta}}_j$. Hence, conditioned on the "good" event that occurs with high probability,

$$\|\widehat{\boldsymbol{\beta}}(\tau) - \boldsymbol{\beta}^*(\tau)\|_\Sigma = \|\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}^*(\tau)\|_\Sigma \leq \|\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}^*_j\|_\Sigma + \|\boldsymbol{\beta}^*(\tau) - \boldsymbol{\beta}^*(\tau_j)\|_\Sigma \leq r_j + \underline{f}^{-1}\delta^*.$$

Taking the maximum over j on the right-hand side, and then the supremum over $\tau \in [\tau_L, \tau_U]$ on the left, we obtain

$$\sup_{\tau \in [\tau_L, \tau_U]} \|\widehat{\boldsymbol{\beta}}(\tau) - \boldsymbol{\beta}^*(\tau)\|_\Sigma \leq \max_{0 \leq j \leq m} r_j + \underline{f}^{-1}\delta^* = r_m + \underline{f}^{-1}\delta^*,$$

completing the proof. □

C.2.3 Proof of Theorem 4.3.2

Similarly to the proof of Theorem 4.3.1, we divide the proof into two stages. In stage one, we prove a uniform bound over the grid points $\{\tau_0, \ldots, \tau_m\}$; in stage two, we prove the claimed bound (4.16) which holds uniformly over the interval $[\tau_L, \tau_U]$.

STAGE ONE. As before, we write $\mathbf{J}_j = \mathbf{J}(\tau_j)$ and $\mathbf{H}_j = \mathbf{H}(\tau_j)$ for $j = 0, \ldots, m$, and define the discretized integrated error up to τ_j as

$$\widetilde{e}_0 := \mathbf{J}_0(\widehat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}^*_0) \text{ and } \widetilde{e}_j := \underbrace{\mathbf{J}_j(\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}^*_j)}_{\text{current step}} + \underbrace{\sum_{\ell=0}^{j-1} w_\ell \mathbf{H}_\ell(\widehat{\boldsymbol{\beta}}_\ell - \boldsymbol{\beta}^*_\ell)}_{\text{preceding steps}}, \quad j = 1, \ldots, m. \tag{C.18}$$

Let $\{r_j\}_{j=0}^m$ be the sequence of radii from the proof of Theorem 4.3.1. We will show that

$$\sup_{j=0,\ldots,m} ||\widetilde{e}_j + Q^*_j||_{\Sigma^{-1}} \lesssim \left(m_4^{1/2} \sqrt{\frac{p + \log n}{nh}} + m_3 r_j + h \right) \cdot r_j, \tag{C.19}$$

holds with probability at least $1 - C_2 n^{-1}$ for some absolute constant $C_2 > 0$, where

$$Q_0^* = \widehat{Q}_0(\boldsymbol{\beta}_0^*) = \frac{1}{n}\sum_{i=1}^{n}\{\Delta_i \bar{K}_h(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0^* - y_i) - \tau_0\}\boldsymbol{x}_i \quad \text{and} \tag{C.20}$$

$$Q_j^* = \frac{1}{n}\sum_{i=1}^{n}\left\{\Delta_i \bar{K}_h(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_j^* - y_i) - \sum_{\ell=0}^{j-1} w_\ell \bar{K}_h(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_\ell^*) - \tau_0\right\}\boldsymbol{x}_i \text{ for } j \in [m]. \tag{C.21}$$

We prove the claim (C.19) in a sequential manner, conditioned on some "good" events. Set

$$\mathscr{A} = \bigcap_{j=0}^{m}\big\{\|\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^*\|_\Sigma \le r_j\big\}, \text{ satisfying } \mathbb{P}(\mathscr{A}) \ge 1 - C_1 n^{-1},$$

For an increasing sequence $0 < \lambda_0 \le \lambda_1 \le \cdots \le \lambda_m$ to be determined, define the event

$$\mathscr{E} = \bigcap_{j=0}^{m-1}\big\{\varpi_j(r_j) \le \lambda_j \cdot r_j, \ \ \omega_j(r_j) \le \lambda_j \cdot r_j\big\}\bigcap\big\{\omega_j(r_m) \le \lambda_m \cdot r_m\big\}, \tag{C.22}$$

where $\varpi_j(\cdot)$'s and $\omega_j(\cdot)$'s are defined in (C.4) and (C.5), respectively. Recall that $\widehat{Q}_0(\widehat{\boldsymbol{\beta}}_0) = \widehat{Q}_1(\widehat{\boldsymbol{\beta}}_1) = \cdots = \widehat{Q}_m(\widehat{\boldsymbol{\beta}}_m) = \mathbf{0}$. Conditioning on the event $\mathscr{A} \cap \mathscr{E}$, we have

$$||\widetilde{e}_0 + Q_0^*||_{\Sigma^{-1}} = \|\widehat{Q}_0(\widehat{\boldsymbol{\beta}}_0) - \widehat{Q}_0(\boldsymbol{\beta}_0^*) - \mathbf{J}_0(\widehat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}_0^*)\|_{\Sigma^{-1}} \le \lambda_0 r_0,$$

and for $j \in [m]$,

$$
\begin{aligned}
||\widetilde{e}_j + Q_j^*||_{\Sigma^{-1}} &= \Bigg\| \mathbf{J}_j(\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^*) - \widehat{Q}_j(\widehat{\boldsymbol{\beta}}_j) + \sum_{\ell=0}^{j-1} w_\ell \mathbf{H}_\ell(\widehat{\boldsymbol{\beta}}_\ell - \boldsymbol{\beta}_\ell^*) \\
&\qquad + \frac{1}{n}\sum_{i=1}^n \left\{ \Delta_i \bar{K}_h(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_j^* - y_i) - \sum_{\ell=0}^{j-1} w_\ell \bar{K}_h(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_\ell^*) - \tau_0 \right\} \boldsymbol{x}_i \Bigg\|_{\Sigma^{-1}} \\
&= \Bigg\| \mathbf{J}_j(\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^*) - \frac{1}{n}\sum_{i=1}^n \Delta_i \{\bar{K}_h(\boldsymbol{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_j - y_i) - \bar{K}_h(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_j^* - y_i)\}\boldsymbol{x}_i \\
&\quad + \sum_{\ell=0}^{j-1} w_\ell \mathbf{H}_\ell(\widehat{\boldsymbol{\beta}}_\ell - \boldsymbol{\beta}_\ell^*) + \sum_{\ell=0}^{j-1} w_\ell \cdot \underbrace{\frac{1}{n}\sum_{i=1}^n \{\bar{K}_h(y_i - \boldsymbol{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_\ell) - \bar{K}_h(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_\ell^*)\}\boldsymbol{x}_i}_{=\widehat{\Delta}_\ell \text{ in (C.12)}} \Bigg\|_{\Sigma^{-1}} \\
&\le \Bigg\| \frac{1}{n}\sum_{i=1}^n \Delta_i \{\bar{K}_h(\boldsymbol{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_j - y_i) - \bar{K}_h(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_j^* - y_i)\}\boldsymbol{x}_i - \mathbf{J}_j(\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^*) \Bigg\|_{\Sigma^{-1}} \\
&\quad + \sum_{\ell=0}^{j-1} w_\ell \|\widehat{\Delta}_\ell + \mathbf{H}_\ell(\widehat{\boldsymbol{\beta}}_\ell - \boldsymbol{\beta}_\ell^*)\|_{\Sigma^{-1}} \\
&\le \omega_j(r_j) + \sum_{\ell=0}^{j-1} w_\ell \varpi_\ell(r_\ell) \le \lambda_j r_j + \sum_{\ell=0}^{j-1} w_\ell \lambda_\ell r_\ell.
\end{aligned}
$$

Recall from the definition of r_j in (C.16), we have $\sum_{\ell=0}^{j-1} w_\ell r_\ell \le (2\overline{f})^{-1}\kappa r_j$, and hence

$$
||\widetilde{e}_j + Q_j^*||_{\Sigma^{-1}} \le \left\{1 + (2\overline{f})^{-1}\kappa\right\} \cdot \lambda_j r_j, \quad j \in [m].
$$

In view of Lemma C.2.5 with $t = 2\log n$, we set

$$
\lambda_j \asymp m_4^{1/2}\sqrt{\frac{p+\log n}{nh}} + m_3 r_j + h, \quad j = 0, 1, \ldots, m,
$$

so that event $\mathscr{E}$ in (C.22) holds with probability at least $1 - 2(m+1)n^{-2}$. This proves (C.19).

STAGE TWO. To generalize (C.19) to (4.16) on the whole process $\widehat{\boldsymbol{\beta}}(\cdot)$, define disjoint intervals $\mathscr{I}_j = [\tau_j, \tau_{j+1})$ for $j = 0, \ldots, m-1$ and $\mathscr{I}_m = \{\tau_m\}$. For any $\tau \in [\tau_L, \tau_U]$, there exists a unique index $j \in \{0, \ldots, m\}$ such that $\tau \in \mathscr{I}_j$ and $\widehat{\boldsymbol{\beta}}(\tau) = \widehat{\boldsymbol{\beta}}_j$. With this notation, the left-hand side of

(4.16) equals

$$\max_{j=0,\ldots,m} \sup_{\tau\in\mathscr{I}_j} \left\| \widehat{\boldsymbol{e}}(\tau) - \frac{1}{n}\sum_{i=1}^{n} \left\{ \tau_L + \int_{\tau_L}^{\tau} \bar{K}_h(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*(u))\mathrm{d}H(u) - \Delta_i \bar{K}_h(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*(\tau) - y_i) \right\} \boldsymbol{x}_i \right\|_{\Sigma^{-1}},$$

where $\widehat{\boldsymbol{e}}(\tau) = \widehat{\boldsymbol{\beta}}_{\mathrm{int}}(\tau) - \boldsymbol{\beta}^*_{\mathrm{int}}(\tau)$. To control the discretization error, conditioning on the event $\mathscr{A}\cap\mathscr{E}$ we have, for each $j = 0, 1, \ldots, m$,

$$\begin{aligned}
&\sup_{\tau\in\mathscr{I}_j} \|\mathbf{J}(\tau)\{\widehat{\boldsymbol{\beta}}(\tau) - \boldsymbol{\beta}^*(\tau)\} - \mathbf{J}_j(\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}^*_j)\|_{\Sigma^{-1}} \\
&\le \sup_{\tau\in\mathscr{I}_j} \|\{\mathbf{J}(\tau) - \mathbf{J}_j\}(\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}^*_j)\|_{\Sigma^{-1}} + \sup_{\tau\in\mathscr{I}_j} \|\mathbf{J}(\tau)\{\boldsymbol{\beta}^*_j - \boldsymbol{\beta}^*(\tau)\}\|_{\Sigma^{-1}} \\
&\le l_1 \underline{f}^{-1} m_3 r_j \delta^* + \overline{g}\underline{f}^{-1}\delta^* \lesssim \delta^*, \qquad \text{(C.23)}
\end{aligned}$$

and

$$\begin{aligned}
&\sup_{\tau\in\mathscr{I}_j} \left\| \int_{\tau_L}^{\tau} \mathbf{H}(u)\{\widehat{\boldsymbol{\beta}}(u) - \boldsymbol{\beta}^*(u)\}\mathrm{d}H(u) - \sum_{\ell=0}^{j-1} \int_{\tau_\ell}^{\tau_{\ell+1}} \mathbf{H}_\ell(\widehat{\boldsymbol{\beta}}_\ell - \boldsymbol{\beta}^*_\ell)\mathrm{d}H(u) \right\|_{\Sigma^{-1}} \qquad \text{(C.24)} \\
&\le \left\| \sum_{\ell=0}^{j-1} \int_{\tau_\ell}^{\tau_{\ell+1}} \left[\mathbf{H}(u)\{\widehat{\boldsymbol{\beta}}(u) - \boldsymbol{\beta}^*(u)\} - \mathbf{H}_\ell(\widehat{\boldsymbol{\beta}}_\ell - \boldsymbol{\beta}^*_\ell)\right]\mathrm{d}H(u) \right\|_{\Sigma^{-1}} \\
&\quad + \sup_{\tau\in\mathscr{I}_j} \left\| \int_{\tau_j}^{\tau} \mathbf{H}(u)\{\widehat{\boldsymbol{\beta}}(u) - \boldsymbol{\beta}^*(u)\}\mathrm{d}H(u) \right\|_{\Sigma^{-1}} \lesssim \log\left(\frac{1-\tau_0}{1-\tau_{j+1}}\right)\cdot\delta^*.
\end{aligned}$$

Next we control the approximation error for discretizing the process $(1/n)\sum_{i=1}^n \boldsymbol{U}_i(\cdot)$. For any interval $\mathscr{I}_j$, write $\boldsymbol{v}(\tau) = \boldsymbol{\beta}^*(\tau) - \boldsymbol{\beta}^*_j$ for $\tau\in\mathscr{I}_j$, satisfying $\|\boldsymbol{v}(\tau)\|_\Sigma \le \underline{f}^{-1}(\tau_{j+1} - \tau_j)$

by the Lipschitz continuity of $\boldsymbol{\beta}^*(\cdot)$. Moreover, we have

$$
\begin{aligned}
&\Big\|\mathbb{E}\Delta_i\big\{\bar{K}_h(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*(\tau)-y_i)-\bar{K}_h(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_j^*-y_i)\big\}\boldsymbol{x}_i\Big\|_{\Sigma^{-1}}\\
&=\sup_{\boldsymbol{u}\in\mathbb{S}^{p-1}}\mathbb{E}\big[\Delta_i\{\bar{K}_h(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*(\tau)-y_i)-\bar{K}_h(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_j^*-y_i)\}\langle\boldsymbol{u},\boldsymbol{z}_i\rangle\big]\\
&=\sup_{\boldsymbol{u}\in\mathbb{S}^{p-1}}\mathbb{E}\int_{-\infty}^{\infty}\left\{\bar{K}\left(\frac{\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*(\tau)-u}{h}\right)-\bar{K}\left(\frac{\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_j^*-u}{h}\right)\right\}g(u|\boldsymbol{x})\mathrm{d}u\cdot\langle\boldsymbol{u},\boldsymbol{z}_i\rangle\\
&=\sup_{\boldsymbol{u}\in\mathbb{S}^{p-1}}h\cdot\mathbb{E}\int_{-\infty}^{\infty}\big\{\bar{K}(v+\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{v}(\tau)/h)-\bar{K}(v)\big\}g(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_j^*-hv|\boldsymbol{x})\mathrm{d}v\cdot\langle\boldsymbol{u},\boldsymbol{z}_i\rangle\\
&\le\overline{g}\sup_{\boldsymbol{u}\in\mathbb{S}^{p-1}}\mathbb{E}\int_{-\infty}^{\infty}\left\{\int_0^1K(v+w\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{v}(\tau)/h)\mathrm{d}w\right\}\mathrm{d}v\cdot\langle\boldsymbol{v}(\tau),\boldsymbol{x}_i\rangle\langle\boldsymbol{u},\boldsymbol{z}_i\rangle\\
&\le\overline{g}\sup_{\boldsymbol{u}\in\mathbb{S}^{p-1}}\mathbb{E}\left(\int_0^1\int_{-\infty}^{\infty}K(v+w\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{v}(\tau)/h)\mathrm{d}v\mathrm{d}w\right)\cdot\langle\boldsymbol{v}(\tau),\boldsymbol{x}_i\rangle\langle\boldsymbol{u},\boldsymbol{z}_i\rangle\\
&\le\overline{g}\sup_{\boldsymbol{u}\in\mathbb{S}^{p-1}}\mathbb{E}|\langle\boldsymbol{v}(\tau),\boldsymbol{x}_i\rangle\langle\boldsymbol{u},\boldsymbol{z}_i\rangle|\le\overline{g}\|\boldsymbol{v}(\tau)\|_{\Sigma}\le\overline{g}\underline{f}^{-1}(\tau_{j+1}-\tau_j).
\end{aligned}
$$

This, combined with Lemma C.2.5–(ii) with $r=\underline{f}^{-1}\delta^*$ and $t=2\log n$, yields that with probability at least $1-(m+1)n^{-2}$,

$$
\begin{aligned}
&\max_{0\le j\le m}\sup_{\tau\in\mathscr{I}_j}\left\|\frac{1}{n}\sum_{i=1}^n\Delta_i\big\{\bar{K}_h(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*(\tau)-y_i)-\bar{K}_h(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_j^*-y_i)\big\}\boldsymbol{x}_i\right\|_{\Sigma^{-1}} \qquad \text{(C.25)}\\
&\quad\lesssim\delta^*+\delta^*m_4^{1/2}\sqrt{\frac{p+\log n}{nh}}\lesssim\delta^*
\end{aligned}
$$

as long as $nh\gtrsim\zeta_p^2(p+\log n)$. Similarly, it follows from Lemma C.2.3–(iii), Lemma C.2.6 and

the union bound that with probability at least $1-2mn^{-2}$,

$$
\begin{aligned}
&\sup_{\tau\in\mathscr{I}_j}\left\|\frac{1}{n}\sum_{i=1}^{n}(1-\mathbb{E})\left\{\int_{\tau_L}^{\tau}\bar{K}_h(y_i-\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*(u))\mathrm{d}H(u)-\sum_{\ell=0}^{j-1}\int_{\tau_\ell}^{\tau_{\ell+1}}\bar{K}_h(y_i-\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_\ell^*)\mathrm{d}H(u)\right\}\boldsymbol{x}_i\right\|_{\Sigma^{-1}}\\
&\le\left\|\frac{1}{n}\sum_{i=1}^{n}\sum_{\ell=0}^{j-1}(1-\mathbb{E})\int_{\tau_\ell}^{\tau_{\ell+1}}\left\{\bar{K}_h(y_i-\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*(u))-\bar{K}_h(y_i-\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_\ell^*)\right\}\mathrm{d}H(u)\cdot\boldsymbol{z}_i\right\|_2\\
&\qquad+\sup_{\tau\in\mathscr{I}_j}\left\|\frac{1}{n}\sum_{i=1}^{n}(1-\mathbb{E})\int_{\tau_j}^{\tau}\bar{K}_h(y_i-\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*(u))\mathrm{d}H(u)\cdot\boldsymbol{z}_i\right\|_2\\
&\le\sum_{\ell=0}^{j-1}\left\|\frac{1}{n}\sum_{i=1}^{n}(1-\mathbb{E})\int_{\tau_\ell}^{\tau_{\ell+1}}\left\{\bar{K}_h(y_i-\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*(u))-\bar{K}_h(y_i-\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_\ell^*)\right\}\mathrm{d}H(u)\cdot\boldsymbol{z}_i\right\|_2\\
&\qquad+\sup_{\tau\in\mathscr{I}_j}\left\|\frac{1}{n}\sum_{i=1}^{n}(1-\mathbb{E})\int_{\tau_j}^{\tau}\bar{K}_h(y_i-\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*(u))\mathrm{d}H(u)\cdot\boldsymbol{z}_i\right\|_2\\
&\lesssim\log\left(\frac{1-\tau_0}{1-\tau_j}\right)\left(m_4^{1/2}\sqrt{\frac{p+\log n}{nh}}+\zeta_p^2\frac{\log n}{nh}\right)\cdot\delta^*\\
&\qquad+\log\left(\frac{1-\tau_j}{1-\tau_{j+1}}\right)\left(\sqrt{\frac{p+\log n}{n}}+\zeta_p\frac{\log n}{n}\right)\cdot\delta^*
\end{aligned}
\tag{C.26}
$$

for all $j=0,1,\ldots,m$. Turning to the deterministic approximation error, it can be shown that for any $j=0,1,\ldots,m-1$ and $\tau\in\mathscr{I}_j$,

$$
\begin{aligned}
&\left\|\mathbb{E}\left\{\int_{\tau_L}^{\tau}\bar{K}_h(y_i-\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*(u))\mathrm{d}H(u)-\sum_{\ell=0}^{j-1}\int_{\tau_\ell}^{\tau_{\ell+1}}\bar{K}_h(y_i-\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_\ell^*)\mathrm{d}H(u)\right\}\boldsymbol{x}_i\right\|_{\Sigma^{-1}}\\
&\qquad\lesssim\log\left(\frac{1-\tau_0}{1-\tau_{j+1}}\right)\cdot\delta^*.
\end{aligned}
$$

Together, (C.19) and (C.23)–(C.26) prove the claimed bounds (4.14)–(4.16).

It remains to control the bias term $\mathbb{E}\boldsymbol{U}_i(\cdot)$. Define the non-smoothed version of $\boldsymbol{U}_i(\cdot)$ as

$$
\boldsymbol{V}_i(\tau)=\left[\tau_L+\int_{\tau_L}^{\tau}\mathbb{1}\{y_i>\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*(u)\}\mathrm{d}H(u)-\Delta_i\mathbb{1}\{y_i<\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*(\tau)\}\right]\boldsymbol{x}_i,\quad\tau\in[\tau_L,\tau_U].
$$

By the martingale property, $\mathbb{E}\boldsymbol{V}_i(\tau) = \mathbf{0}$ for every $\tau \in [\tau_L, \tau_U]$. Note that

$$\boldsymbol{U}_i(\tau) - \boldsymbol{V}_i(\tau) = \Bigg(\Delta_i \big[\mathbb{1}\{y_i < \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}^*(\tau)\} - \bar{K}_h(\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}^*(\tau) - y_i) \big] + \int_{\tau_L}^{\tau} \big[\bar{K}_h(y_i - \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}^*(u)) - \mathbb{1}\{y_i > \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}^*(u)\} \big] \mathrm{d}H(u) \Bigg) \boldsymbol{x}_i.$$

Following the same calculations that lead to (C.67), we obtain

$$\sup_{\tau \in [\tau_L, \tau_U]} \|\mathbb{E}\boldsymbol{U}_i(\tau)\|_{\Sigma^{-1}} = \sup_{\tau \in [\tau_L, \tau_U]} \|\mathbb{E}\{\boldsymbol{U}_i(\tau) - \boldsymbol{V}_i(\tau)\}\|_{\Sigma^{-1}} \leq 0.5 l_1 \kappa_2 \big\{ 1 + \log\big(\tfrac{1-\tau_L}{1-\tau_U}\big) \big\} h^2.$$

This completes the proof of the theorem. □

C.2.4 Proof of Theorem 4.3.3

Assume without loss of generality that $||\boldsymbol{a}_n||_\Sigma = 1$; otherwise, we simply replace $\boldsymbol{a}_n$ by $\boldsymbol{a}_n / ||\boldsymbol{a}_n||_\Sigma$. Following the general result in Theorem 1.5.4 of van der Vaart and Wellner [1996], the claimed weak convergence (4.22) is a direct consequence of the weak convergence of finite-dimensional marginals and the asymptotic tightness of $\mathbb{G}_n(\cdot)$.

For the former, via the Cramér–Wold device, it is equivalent to show that for any finite set of values $\{\tau_\ell\}_{\ell=1}^L \subseteq [\tau_L, \tau_U]$ and $(\gamma_1, \ldots, \gamma_L)^{\mathrm{T}} \in \mathbb{R}^L$,

$$\sum_{\ell=1}^{L} \gamma_\ell \mathbb{G}_n(\tau_\ell) \xrightarrow{\mathrm{d}} \sum_{\ell=1}^{L} \gamma_\ell \mathbb{G}(\tau_\ell), \tag{C.27}$$

with $\mathbb{G}(\cdot)$ defined in (4.22). For $i = 1, \ldots, n$, define centered variables $W_i = \sum_{\ell=1}^L \gamma_\ell \langle \boldsymbol{a}_n, \boldsymbol{U}_{0i}(\tau_\ell) \rangle$

with $\boldsymbol{U}_{0i}(\tau) := \boldsymbol{U}_i(\tau) - \mathbb{E}\boldsymbol{U}_i(\tau)$, so that $\sum_{\ell=1}^{L} \gamma_\ell \mathbb{G}_n(\tau_\ell) = n^{-1/2}\sum_{i=1}^{n} W_i$. Moreover,

$$
\begin{aligned}
\mathrm{Var}(W_i) &= \mathrm{Var}\left(\sum_{\ell=1}^{L} \gamma_\ell \langle \boldsymbol{a}_n, \boldsymbol{U}_{0i}(\tau_\ell) \rangle \right) \\
&= \sum_{\ell_1=1}^{L} \sum_{\ell_2=1}^{L} \gamma_{\ell_1}\gamma_{\ell_2} \cdot \mathrm{Cov}\left(\langle \boldsymbol{a}_n, \boldsymbol{U}_{0i}(\tau_{\ell_1}) \rangle, \langle \boldsymbol{a}_n, \boldsymbol{U}_{0i}(\tau_{\ell_2}) \rangle \right) \\
&= \sum_{\ell_1=1}^{L} \sum_{\ell_2=1}^{L} \gamma_{\ell_1}\gamma_{\ell_2} \cdot \boldsymbol{a}_n^{\mathrm{T}} \mathbb{E}\big\{ \boldsymbol{U}_{0i}(\tau_{\ell_1}) \boldsymbol{U}_{0i}(\tau_{\ell_2})^{\mathrm{T}} \big\} \boldsymbol{a}_n \\
&= \sum_{\ell_1=1}^{L} \sum_{\ell_2=1}^{L} \gamma_{\ell_1}\gamma_{\ell_2} \cdot \boldsymbol{a}_n^{\mathrm{T}} \mathbb{E}\big\{ \boldsymbol{U}_{i}(\tau_{\ell_1}) \boldsymbol{U}_{i}(\tau_{\ell_2})^{\mathrm{T}} \big\} \boldsymbol{a}_n \\
&\quad - \sum_{\ell_1=1}^{L} \sum_{\ell_2=1}^{L} \gamma_{\ell_1}\gamma_{\ell_2} \cdot \boldsymbol{a}_n^{\mathrm{T}} \mathbb{E}\big\{ \boldsymbol{U}_{i}(\tau_{\ell_1}) \big\} \mathbb{E}\big\{ \boldsymbol{U}_{i}(\tau_{\ell_2})^{\mathrm{T}} \big\} \boldsymbol{a}_n \\
&\to \sum_{\ell_1=1}^{L} \sum_{\ell_2=1}^{L} \gamma_{\ell_1}\gamma_{\ell_2} \cdot H(\ell_1, \ell_2) = \mathrm{Var}\left(\sum_{\ell=1}^{L} \gamma_\ell \mathbb{G}(\tau_\ell) \right) \text{ as } n \to \infty,
\end{aligned}
$$

where $H(\cdot,\cdot)$ is defined in (4.21). The finite-dimensional weak convergence (C.27) then follows from the central limit theorem.

Turning to the asymptotic tightness of $\mathbb{G}_n(\cdot)$, an equivalent characterization is the asymptotic uniform equicontinuity in probability; see Theorem 1.5.7 in van der Vaart and Wellner [1996] and the definition above it. That is, for any $x > 0$,

$$
\lim_{\delta \to 0} \limsup_{n \to \infty} \mathbb{P}\left\{ \sup_{|\tau_1 - \tau_2| < \delta} \left| \mathbb{G}_n(\tau_1) - \mathbb{G}_n(\tau_2) \right| > x \right\} = 0,
$$

which is ensured by Lemma C.2.8.

Finally, the existence of almost surely continuous sample paths of $\mathbb{G}(\cdot)$ follows from Addendum 1.5.8 in van der Vaart and Wellner [1996]. □

C.3 Proofs of the Main Results in Section 4.3.3

C.3.1 Technical lemmas

In this section, we provide the technical lemmas needed to establish the validity of the multiplier bootstrap procedure. Recall that e_i's are i.i.d. Rademacher random variables that are independent of the observed data $\mathbb{D}_n = \{y_i, \Delta_i, \boldsymbol{x}_i\}_{i=1}^n$. Similarly to (C.3), we define the symmetrized Bregman divergence $D^\flat : \mathbb{R}^p \times \mathbb{R}^p \to [0,\infty)$ in the bootstrap world as

$$D^\flat(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = \langle \widehat{Q}_j^\flat(\boldsymbol{\beta}_1) - \widehat{Q}_j^\flat(\boldsymbol{\beta}_2), \boldsymbol{\beta}_1 - \boldsymbol{\beta}_2 \rangle, \tag{C.28}$$

which is also independent of j, where $\widehat{Q}_j^\flat(\cdot)$'s are the randomly perturbed estimating equations defined in (4.7) and (4.8).

Lemma C.3.1 (Conditional restricted strong convexity)**.** *Assume Conditions 4.3.1 and 4.3.3 hold. Let $r = h/(4\eta_{1/4})$ with $\eta_{1/4}$ defined in* (C.2)*, and $t \geq 0$. Suppose the "effective" sample size satisfies $nh \gtrsim \max\{p, \zeta_p t^{1/2}\}$. Then, there exists an event $\mathscr{E}_1(t)$ with $\mathbb{P}\{\mathscr{E}_1(t)\} \geq 1-(m+3)e^{-t}$ such that, conditioning on $\mathscr{E}_1(t)$,*

$$\inf_{j\in\{0,\ldots,m\}} \inf_{\boldsymbol{\beta}\in\boldsymbol{\beta}_j^*+\Theta(r)} \frac{D(\boldsymbol{\beta}, \boldsymbol{\beta}_j^*)}{\kappa_l \|\boldsymbol{\beta}-\boldsymbol{\beta}_j^*\|_\Sigma^2} \geq \frac{1}{2}\underline{g}, \text{ and}$$
$$\mathbb{P}^*\left\{ \inf_{j\in\{0,\ldots,m\}} \inf_{\boldsymbol{\beta}\in\boldsymbol{\beta}_j^*+\Theta(r)} \frac{D^\flat(\boldsymbol{\beta}, \boldsymbol{\beta}_j^*)}{\kappa_l \|\boldsymbol{\beta}-\boldsymbol{\beta}_j^*\|_\Sigma^2} \geq \frac{1}{2}\underline{g} \right\} \geq 1-(m+1)e^{-t}.$$

Lemma C.3.2. *Assume Condition 4.3.2 holds. Let $\{\xi_i\}_{i=1}^n$ be independent random variables satisfying $|\xi_i| \leq M$ for some $M > 0$, and $\{e_i\}_{i=1}^n$ are Rademacher random variables independent of the data $\{\boldsymbol{x}_i, \xi_i\}_{i=1}^n$. Then, there exists an event $\mathscr{E}_2$ depending on $\{\boldsymbol{x}_i, \xi_i\}_{i=1}^n$ such that (i) $\mathbb{P}(\mathscr{E}_2) \geq 1-n^{-2}$, and (ii) conditioned on $\mathscr{E}_2$,*

$$\left\| \frac{1}{n}\sum_{i=1}^n (e_i \cdot \xi_i \boldsymbol{z}_i) \right\|_2 \lesssim \sqrt{\frac{p+\log n}{n}}$$

holds with $\mathbb{P}^*$-probability (over $\{e_i\}_{i=1}^n$) at least $1-n^{-2}$ as long as the sample size satisfies $n \gtrsim \zeta_p^2 \log n$.

The following lemma provides upper bounds for two Rademacher weighted stochastic processes. For $j = 0, 1, \ldots, m$ and any $r > 0$, define

$$\Gamma_j(r) := \sup_{\boldsymbol{\beta} \in \boldsymbol{\beta}_j^* + \Theta(r)} \left\| \frac{1}{n} \sum_{i=1}^n e_i \big\{ \bar{K}_h(y_i - \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}) - \bar{K}_h(y_i - \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_j^*) \big\} \boldsymbol{z}_i \right\|_2, \tag{C.29}$$

$$\Gamma_j^{\Delta}(r) := \sup_{\boldsymbol{\beta} \in \boldsymbol{\beta}_j^* + \Theta(r)} \left\| \frac{1}{n} \sum_{i=1}^n e_i \Delta_i \big\{ \bar{K}_h(\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta} - y_i) - \bar{K}_h(\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_j^* - y_i) \big\} \boldsymbol{z}_i \right\|_2, \tag{C.30}$$

where $\{e_i\}_{i=1}^n$ is a sequence of independent Rademacher random variables that are independent of $\{y_i, \Delta_i, \boldsymbol{x}_i\}_{i=1}^n$, and $\boldsymbol{z}_i = \Sigma^{-1/2} \boldsymbol{x}_i$.

Lemma C.3.3. *Assume that Conditions 4.3.1–4.3.3 hold, and $K(\cdot)$ in Condition 4.3.1 is l_K-Lipschitz continuous. Given any $0 < r \le \zeta_p$, there exists an event $\mathscr{E}_3$ with $\mathbb{P}(\mathscr{E}_3) \ge 1 - 3n^{-1}$ such that, with $\mathbb{P}^*$-probability at least $1-(m+1)n^{-2}$ conditioned on $\mathscr{E}_3$,*

$$\sup_{j \in \{0,\ldots,m\}} \Gamma_j(r) \lesssim r \sqrt{\frac{p + \log n}{nh}} \left(m_4^{1/2} + \zeta_p^2 \sqrt{\frac{p \log n}{nh}} \right),$$

provided $n \gtrsim \zeta_p^2 \log n$. The same uniform bound also applies to $\Gamma_j^{\Delta}(r)$.

Lemma C.3.4. *Assume that Conditions 4.3.1–4.3.3 hold, and $K(\cdot)$ in Condition 4.3.1 is l_K-Lipschitz continuous. Then, there exists an event $\mathscr{E}_4$ with $\mathbb{P}(\mathscr{E}_4) \ge 1 - (m+3n+1)n^{-2}$ such that, with $\mathbb{P}^*$-probability at least $1-n^{-2}$ conditional on $\mathscr{E}_4$,*

$$\begin{aligned} &\sup_{\boldsymbol{\beta} \in \boldsymbol{\beta}_j^* + \Theta(r)} \| \mathbf{J}_j(\boldsymbol{\beta} - \boldsymbol{\beta}_j^*) - \{ \widehat{Q}_j^{\natural}(\boldsymbol{\beta}) - \widehat{Q}_j^{\natural}(\boldsymbol{\beta}_j^*) \} \|_{\Sigma^{-1}} \\ &\quad \lesssim \left\{ m_4^{1/2} \sqrt{\frac{p + \log n}{nh}} + m_3 r + h + \zeta_p^2 \frac{(p \log n)^{1/2} (p + \log n)^{1/2}}{nh} \right\} \cdot r \end{aligned}$$

holds uniformly over $j = 0, 1, \ldots, m$, where $\mathbf{J}_j = \mathbb{E}\{ g(\boldsymbol{x}^{\mathrm{T}} \boldsymbol{\beta}_j^* | \boldsymbol{x}) \boldsymbol{x} \boldsymbol{x}^{\mathrm{T}} \}$.

Lemma C.3.5. *Assume Conditions 4.3.1–4.3.4 hold, and let $r > 0$. Then, there exists an event $\mathscr{E}_5$ with $\mathbb{P}(\mathscr{E}_5) \geq 1 - mn^{-2}$ such that conditioning on $\mathscr{E}_5$,*

$$\sup_{\cap_{\ell=0}^{j-1}\{\boldsymbol{\beta}_\ell \in \boldsymbol{\beta}_\ell^* + \Theta(r)\}} \left\| \frac{1}{n}\sum_{i=1}^{n}\sum_{\ell=0}^{j-1}\int_{\tau_\ell}^{\tau_{\ell+1}} \mathrm{d}H(u)\Big[\mathbf{H}_\ell(\boldsymbol{\beta}_\ell - \boldsymbol{\beta}_\ell^*) - W_i\big\{\bar{K}_h(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_\ell^*) - \bar{K}_h(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_\ell)\big\}\boldsymbol{x}_i\Big] \right\|_{\Sigma^{-1}}$$
$$\lesssim \log\big(\tfrac{1-\tau_0}{1-\tau_j}\big) \cdot \left\{ \left(m_4^{1/2}\sqrt{\frac{p+\log n}{nh}} + m_3 r + h \right) \cdot r + \max_{\ell=0,\ldots,j-1} \Gamma_\ell(r) \right\}$$

holds for all $j = 1, \ldots, m$, where $\mathbf{H}_\ell = \mathbb{E}\{f(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_\ell^* | \boldsymbol{x})\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\}$ and $\Gamma_\ell(r)$ is defined in (C.29).

The following lemma establishes the asymptotic uniform equicontinuity of the process $\mathbb{G}_n^\flat(\cdot) = n^{-1/2}\sum_{i=1}^n e_i\langle \boldsymbol{a}_n, \boldsymbol{U}_i(\cdot)\rangle$ in $\ell^\infty([\tau_L, \tau_U])$, thus validating the asymptotic tightness of $\mathbb{G}_n^\flat(\cdot)$, where $\boldsymbol{U}_i(\cdot)$ is defined in (4.15).

Lemma C.3.6. *Assume that the conditions of Theorem 4.3.6 hold. For any $x > 0$ and sequence of vectors $\boldsymbol{a}_n$ satisfying $\|\boldsymbol{a}_n\|_\Sigma = 1$, conditioned on any observed data $\mathbb{D}_n = \{(y_i, \Delta_i, \boldsymbol{x}_i\}_{i=1}^n$, we have*

$$\lim_{\delta \to 0} \limsup_{n \to \infty} \mathbb{P}^* \left\{ \sup_{|\tau_1 - \tau_2| < \delta} \left| \mathbb{G}_n^\flat(\tau_1) - \mathbb{G}_n^\flat(\tau_2) \right| > x \right\} = 0. \tag{C.31}$$

where $\mathbb{G}_n^\flat(\cdot) = n^{-1/2}\sum_{i=1}^n e_i\langle \boldsymbol{a}_n, \boldsymbol{U}_i(\cdot)\rangle$ with $\boldsymbol{U}_i(\tau)$ defined in (4.15).

C.3.2 Proof of Theorem 4.3.4

Similar to the proof of Theorem 4.3.1, we first prove a uniform bound over the grid points $\{\tau_0, \ldots, \tau_m\}$. Recall the bootstrapped SEE $\widehat{Q}_j^\flat(\boldsymbol{\beta})$ given in (4.7) and (4.8), and $\widehat{Q}_j^\flat(\widehat{\boldsymbol{\beta}}_j^\flat) = \mathbf{0}$. Following the localized argument as in the proof of Theorem 4.3.1, for the same radius parameter $r^\diamondsuit$ therein, define $\widetilde{\boldsymbol{\beta}}_j^\flat = \boldsymbol{\beta}_j^* + \gamma_j(\widehat{\boldsymbol{\beta}}_j^\flat - \boldsymbol{\beta}_j^*)$ with $\gamma_j := \sup\big\{\gamma \in [0,1] : \gamma(\widehat{\boldsymbol{\beta}}_j^\flat - \boldsymbol{\beta}_j^*) \in \Theta(r^\diamondsuit)\big\}$, so that $\widetilde{\boldsymbol{\beta}}_j^\flat = \widehat{\boldsymbol{\beta}}_j^\flat$ if $\widehat{\boldsymbol{\beta}}_j^\flat \in \boldsymbol{\beta}_j^* + \Theta(r^\diamondsuit)$ and $\widetilde{\boldsymbol{\beta}}_j^\flat \in \boldsymbol{\beta}_j^* + \partial\Theta(r^\diamondsuit)$ if $\widehat{\boldsymbol{\beta}}_j^\flat \notin \boldsymbol{\beta}_j^* + \Theta(r^\diamondsuit)$. Consequently,

$$D^\flat(\widetilde{\boldsymbol{\beta}}_j^\flat, \boldsymbol{\beta}_j^*) \leq \rho_j \cdot \langle -\widehat{Q}_j^\flat(\boldsymbol{\beta}_j^*), \widehat{\boldsymbol{\beta}}_j^\flat - \boldsymbol{\beta}_j^*\rangle \leq \|\widehat{Q}_j^\flat(\boldsymbol{\beta}_j^*)\|_{\Sigma^{-1}} \cdot \|\widetilde{\boldsymbol{\beta}}_j^\flat - \boldsymbol{\beta}_j^*\|_\Sigma,$$

where $D^\flat$ is defined in (C.28).

In addition to $\mathscr{F}$ in (C.10), define

$$\mathscr{F}^\flat = \bigcap_{j=0}^{m} \left\{ D^\flat(\boldsymbol{\beta}, \boldsymbol{\beta}_j^*) \geq \kappa \cdot \|\boldsymbol{\beta} - \boldsymbol{\beta}_j^*\|_\Sigma^2 \text{ for all } \boldsymbol{\beta} \in \boldsymbol{\beta}_j^* + \Theta(r^\diamondsuit) \right\}. \tag{C.32}$$

Conditioned on $\mathscr{F}^\flat$, we have for all $j = 0, 1, \ldots, m$ that

$$\|\widetilde{\boldsymbol{\beta}}_j^\flat - \boldsymbol{\beta}_j^*\|_\Sigma \leq \kappa^{-1} \|\widehat{Q}_j^\flat(\boldsymbol{\beta}_j^*)\|_{\Sigma^{-1}}.$$

For the bootstrapped estimating function, by the triangle inequality we have

$$\begin{aligned}
&\|\widehat{Q}_j^\flat(\boldsymbol{\beta}_j^*)\|_{\Sigma^{-1}} \\
\leq &\left\| \frac{1}{n} \sum_{i=1}^n e_i \left\{ \Delta_i \bar{K}_h(\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_j^* - y_i) - \tau_0 - \sum_{\ell=0}^{j-1} w_\ell \bar{K}_h(y_i - \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_\ell^*) \right\} \boldsymbol{x}_i \right\|_{\Sigma^{-1}} \\
&+ \left\| \frac{1}{n} \sum_{i=1}^n \sum_{\ell=0}^{j-1} e_i w_\ell \big\{ \bar{K}_h(y_i - \boldsymbol{x}_i^{\mathrm{T}} \widehat{\boldsymbol{\beta}}_\ell^\flat) - \bar{K}_h(y_i - \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_\ell^*) \big\} \boldsymbol{x}_i \right\|_{\Sigma^{-1}} \\
&+ \left\| \widehat{Q}_0(\boldsymbol{\beta}_j^*) - Q_0(\boldsymbol{\beta}_j^*) - \sum_{\ell=0}^{j-1} w_\ell \big(\widehat{\Delta}_\ell^\flat + \Delta_\ell \big) + Q_0(\boldsymbol{\beta}_j^*) - \sum_{\ell=0}^{j-1} w_\ell \mathbb{E}\{ \bar{K}_h(y - \boldsymbol{x}^{\mathrm{T}} \boldsymbol{\beta}_\ell^*) \boldsymbol{x}_i \} \right\|_{\Sigma^{-1}} \\
\leq &\Bigg\| \underbrace{\frac{1}{n} \sum_{i=1}^n e_i \left\{ \Delta_i \bar{K}_h(\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_j^* - y_i) - \tau_0 - \sum_{\ell=0}^{j-1} w_\ell \bar{K}_h(y_i - \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_\ell^*) \right\} \boldsymbol{x}_i}_{:= \widetilde{Q}_j^\flat(\boldsymbol{\beta}_j^*)} \Bigg\|_{\Sigma^{-1}} \\
&+ \|\widehat{Q}_0(\boldsymbol{\beta}_j^*) - Q_0(\boldsymbol{\beta}_j^*)\|_{\Sigma^{-1}} + \sum_{\ell=0}^{j-1} w_\ell \big(\|\widetilde{\Delta}_\ell^\flat\|_{\Sigma^{-1}} + \|\widehat{\Delta}_\ell^\flat\|_{\Sigma^{-1}} + \|\Delta_\ell\|_{\Sigma^{-1}} \big) \\
&+ \left\| Q_0(\boldsymbol{\beta}_j^*) - \sum_{\ell=0}^{j-1} w_\ell \mathbb{E}\{ \bar{K}_h(y - \boldsymbol{x}^{\mathrm{T}} \boldsymbol{\beta}_\ell^*) \boldsymbol{x}_i \} \right\|_{\Sigma^{-1}},
\end{aligned} \tag{C.33}$$

for $j \geq 1$, where $w_\ell = H(\tau_{\ell+1}) - H(\tau_\ell)$,

$$\widetilde{\Delta}_\ell^\flat = \frac{1}{n}\sum_{i=1}^n e_i\big\{\bar{K}_h(y_i - \boldsymbol{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_\ell^\flat) - \bar{K}_h(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_\ell^*)\big\}\boldsymbol{x}_i,$$
$$\widehat{\Delta}_\ell^\flat = \frac{1}{n}\sum_{i=1}^n \big\{\bar{K}_h(y_i - \boldsymbol{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_\ell^\flat) - \bar{K}_h(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_\ell^*)\big\}\boldsymbol{x}_i$$

and Δ_ℓ is defined in (C.12). In particular,

$$\|\widehat{Q}_0^\flat(\boldsymbol{\beta}_0^*)\|_{\Sigma^{-1}} = \left\|\frac{1}{n}\sum_{i=1}^n e_i\left\{\Delta_i\bar{K}_h(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0^* - y_i) - \tau_0\right\}\boldsymbol{x}_i\right\|_{\Sigma^{-1}} + \|\widehat{Q}_0(\boldsymbol{\beta}_0^*) - Q_0(\boldsymbol{\beta}_0^*)\|_{\Sigma^{-1}} + \|Q_0(\boldsymbol{\beta}_0^*)\|_{\Sigma^{-1}}.$$

The rest of the proof is similar to that of Theorem 4.3.1, and thus we skip some of the technical details. For some $\delta > 0$ to be determined, define the event

$$\mathscr{G}^\flat = \left\{\max_{0\leq j\leq m}\|\widetilde{Q}_j^\flat(\boldsymbol{\beta}_j^*)\|_{\Sigma^{-1}} \bigvee \max_{0\leq j\leq m}\|\widehat{Q}_0(\boldsymbol{\beta}_j^*) - Q_0(\boldsymbol{\beta}_j^*)\|_{\Sigma^{-1}} \bigvee \max_{0\leq \ell\leq m-1}\|\Delta_\ell\|_{\Sigma^{-1}} \leq \delta\right\}. \tag{C.34}$$

Conditioned on $\mathscr{F}^\flat \cap \mathscr{G}^\flat$, it follows from Lemma C.2.7, (C.33) and (C.34) that

$$\|\widehat{Q}_0^\flat(\boldsymbol{\beta}_0^*)\|_{\Sigma^{-1}} < 2\delta + a,$$
$$\widehat{Q}_j^\flat(\boldsymbol{\beta}_j^*)\|_{\Sigma^{-1}} < 2\delta + a + \sum_{\ell=0}^{j-1} w_\ell\big(\delta + a + \|\widetilde{\Delta}_\ell^\flat\|_{\Sigma^{-1}} + \|\widehat{\Delta}_\ell\|_{\Sigma^{-1}}\big), \quad j = 1,\ldots,m,$$

where a is defined in (C.14). Similarly to (C.17), conditioned on the "good" event

$$\mathscr{F}^\flat \cap \mathscr{G}^\flat \cap \bigcap_{\ell=0}^{m-1}\big\{\varpi_\ell(r_\ell) \vee \Gamma_\ell(r_\ell) \leq \overline{f} r_\ell\big\}, \tag{C.35}$$

the convergence radii $\{r_j\}_{j=0}^m$ are recursively defined as

$$r_j = \big(1+3\kappa^{-1}\overline{f}w_{j-1}\big)r_{j-1} + \kappa^{-1}w_{j-1}(\delta+a),\ \ j=1,\ldots,m,\ \text{ and }\ r_0 = \kappa^{-1}(2\delta+a). \tag{C.36}$$

Denoting $C = \kappa^{-1}(3\overline{f}+1)$, it follows that $r_j \le \big(\frac{1-\tau_L}{1-\tau_j}\big)^C \cdot r_0$.

Next we complement the above deterministic analysis with probabilistic bounds. By Lemmas C.2.3, C.2.5 and Lemmas C.3.1–C.3.3, we choose $\kappa = (\underline{g}\kappa_l)/2$ and $\delta \asymp \sqrt{(p+\log n)/n} + \zeta_p \log(n)/n$ so that there exists an event $\mathscr{E}$ with $\mathbb{P}(\mathscr{E}) \ge 1 - C_1 n^{-1}$ such that conditioned on $\mathscr{E}$,

$$\mathbb{P}^*(\mathscr{F}^\flat \cap \mathscr{G}^\flat) \ge 1 - C_2 n^{-1},$$
$$\sup_{\ell\in\{0,\ldots,m-1\}} \varpi_\ell(r_\ell) \lesssim \left(m_4^{1/2}\sqrt{\frac{p+\log n}{nh}} + m_3 r_\ell + h \right)\cdot r_\ell$$

and

$$\sup_{\ell\in\{0,\ldots,m-1\}} \Gamma_\ell(r_\ell) \lesssim r_\ell \sqrt{\frac{p+\log n}{nh}} \left(m_4^{1/2} + \zeta_p^2 \sqrt{\frac{p\log n}{nh}} \right)$$

with probability at least $1-n^{-1}$, provided $nh \gtrsim \zeta_p^2 (p+\log n)^{1/2}$ and $n \gtrsim \zeta_p^2 \log n$. Moreover, the uniform bound (4.11) holds conditioned on $\mathscr{E}$. Consequently, it follows from (C.36) and (C.14) that

$$r_j \le \left(\frac{1-\tau_L}{1-\tau_j}\right)^C \cdot r_0 \asymp \left(\frac{1-\tau_L}{1-\tau_j}\right)^C \underline{g}^{-1} \left(\sqrt{\frac{p+\log n}{n}} + \zeta_p \frac{\log n}{n} + h^2 \right)$$

holds uniformly over $j \in \{0,\ldots,m\}$.

Recall that m_3 and m_4 are dimension-free moment parameters. Given the bandwidth $h = h_n \asymp \{(p+\log n)/n\}^\gamma$ with $\gamma \in [1/4, 1/2)$, and under the sample size requirement $n \gtrsim \zeta_p^{2/(1-\gamma)} (p+\log n)^{(1/2-\gamma)/(1-\gamma)} (p\log n)^{1/(2-2\gamma)}$, we conclude that conditioned on $\mathscr{E}$, the "good"

event (C.35) occurs with $\mathbb{P}^*$-probability at least $1-C_3n^{-1}$, and

$$\left(\frac{1-\tau_L}{1-\tau_j}\right)^C \underline{g}^{-1}\sqrt{\frac{p+\log n}{n}} \asymp r_j \le r^{\diamondsuit} \asymp \left(\frac{p+\log n}{n}\right)^{\gamma} \text{ for all } j=0,1,\ldots,m.$$

This proves the uniform bound over $\tau \in \{\tau_0,\tau_1,\ldots,\tau_m\}$, which can naturally be extended to $\tau \in [\tau_L,\tau_U]$ following the last paragraph in the proof of Theorem 4.3.1. □

C.3.3 Proof of Theorem 4.3.5

We divide the proof into two steps as in the proof of Theorem 4.3.2. Recall that $W_i = e_i + 1$, where e_i's are independent Rademacher variables.

STEP 1. (Uniform bound over $\{\tau_0,\ldots,\tau_m\}$) For simplicity, we write $\mathbf{J}_j = \mathbf{J}(\tau_j)$ and $\mathbf{H}_j = \mathbf{H}(\tau_j)$ for $j=0,\ldots,m$, and define the accumulated bootstrap errors as

$$\begin{aligned}
&\widetilde{e}^{\flat}_{\text{int}}(\tau_0) := \mathbf{J}_0(\widehat{\boldsymbol{\beta}}^{\flat}_0 - \widehat{\boldsymbol{\beta}}_0) \text{ and} \\
&\widetilde{e}^{\flat}_{\text{int}}(\tau_j) := \mathbf{J}_j(\widehat{\boldsymbol{\beta}}^{\flat}_j - \widehat{\boldsymbol{\beta}}_j) + \sum_{\ell=0}^{j-1}\int_{\tau_\ell}^{\tau_{\ell+1}} \mathbf{H}_\ell(\widehat{\boldsymbol{\beta}}^{\flat}_\ell - \widehat{\boldsymbol{\beta}}_\ell)\mathrm{d}H(u), \quad j=1,\ldots,m.
\end{aligned}$$

We claim that there exits an event $\mathscr{F}$ on which (4.14)–(4.16) hold such that $\mathbb{P}(\mathscr{F}) \ge 1-C_4n^{-1}$, and

$$\sup_{j=0,\ldots,m} ||\widetilde{e}^{\flat}_{\text{int}}(\tau_j) + Q^{*\flat}_j||_{\Sigma^{-1}} \lesssim m_4^{1/2}\frac{p+\log n}{nh^{1/2}} + h\sqrt{\frac{p+\log n}{n}} + \zeta_p^2\frac{(p+\log n)(p\log n)^{1/2}}{n^{3/2}h} \tag{C.37}$$

with $\mathbb{P}^*$-probability at least $1-C_5n^{-1}$ conditioned on $\mathscr{F}$, where

$$\begin{aligned}
Q^{*\flat}_0 &= \frac{1}{n}\sum_{i=1}^n e_i\big\{\Delta_i\bar{K}_h(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*_0 - y_i) - \tau_0\big\}\boldsymbol{x}_i, \\
Q^{*\flat}_j &= \frac{1}{n}\sum_{i=1}^n e_i\left\{\Delta_i\bar{K}_h(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*_j - y_i) - \sum_{\ell=0}^{j-1}\int_{\tau_\ell}^{\tau_{\ell+1}} \bar{K}_h(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*_\ell)\,\mathrm{d}H(u) - \tau_0\right\}\boldsymbol{x}_i
\end{aligned}$$

for $j = 1, \ldots, m$.

From Theorem 4.3.4 and its proof, we see that there exists an event $\mathscr{E}_1$ with $\mathbb{P}(\mathscr{E}_1) \geq 1 - C_1 n^{-1}$ such that conditioned on $\mathscr{E}_1$,

$$\max_{0 \leq j \leq m} \|\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^*\|_\Sigma \lesssim \sqrt{\frac{p + \log n}{n}}, \text{ and}$$
$$\mathbb{P}^*\left(\max_{0 \leq j \leq m} \|\widehat{\boldsymbol{\beta}}_j^\flat - \boldsymbol{\beta}_j^*\|_\Sigma \lesssim \sqrt{\frac{p + \log n}{n}} \right) \geq 1 - C_3 n^{-1}. \tag{C.38}$$

We then prove the claim (C.37). For $j = 0$, by the triangle inequality we have

$$\begin{aligned}
&||\widetilde{e}_{\text{int}}^\flat(\tau_0) + Q_0^{*\flat}||_{\Sigma^{-1}} \\
&\leq ||\widetilde{e}_{\text{int}}(\tau_0) + Q_0^*||_{\Sigma^{-1}} + \left\| \mathbf{J}_0(\widehat{\boldsymbol{\beta}}_0^\flat - \boldsymbol{\beta}_0^*) + \frac{1}{n}\sum_{i=1}^n W_i\{\Delta_i \bar{K}_h(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0^* - y_i) - \tau_0\}\boldsymbol{x}_i \right\|_{\Sigma^{-1}} \\
&=: \mathrm{I}_0 + \mathrm{II}_0,
\end{aligned}$$

and for $j \geq 1$,

$$\begin{aligned}
&||\widetilde{e}_{\text{int}}^\flat(\tau_j) + Q_j^{*\flat}||_{\Sigma^{-1}} \\
&\leq ||\widetilde{e}_{\text{int}}(\tau_j) + Q_j^*||_{\Sigma^{-1}} + \left\| \mathbf{J}_j(\widehat{\boldsymbol{\beta}}_j^\flat - \boldsymbol{\beta}_j^*) + \sum_{\ell=0}^{j-1} \int_{\tau_\ell}^{\tau_{\ell+1}} \mathbf{H}_\ell(\widehat{\boldsymbol{\beta}}_\ell^\flat - \boldsymbol{\beta}_\ell^*)\mathrm{d}H(u) \right. \\
&\qquad \left. + \frac{1}{n}\sum_{i=1}^n W_i \left\{ \Delta_i \bar{K}_h(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_j^* - y_i) - \sum_{\ell=0}^{j-1} \int_{\tau_\ell}^{\tau_{\ell+1}} \bar{K}_h(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_\ell^*)\mathrm{d}H(u) - \tau_0 \right\} \boldsymbol{x}_i \right\|_{\Sigma^{-1}} \\
&=: \mathrm{I}_j + \mathrm{II}_j,
\end{aligned}$$

where $\widetilde{e}_{\text{int}}(\tau_j)$ and Q_j^* are defined in (C.18) and (C.20)–(C.21). Let $\mathscr{E}_2$ be the event that (4.14)–(4.16) hold. Then $\mathbb{P}(\mathscr{E}_2) \geq 1 - C_2 n^{-1}$ for some constant C_2, and conditioned on $\mathscr{E}_2$,

$$\max_{0 \leq j \leq m} \mathrm{I}_j = ||\widetilde{e}_{\text{int}}(\tau_j) + Q_j^*||_{\Sigma^{-1}} \lesssim m_4^{1/2} \frac{p + \log n}{n h^{1/2}} + h\sqrt{\frac{p + \log n}{n}}. \tag{C.39}$$

It remains to bound II_j for $j = 0, 1, \ldots, m$. Recall that $\widehat{Q}_j^\flat(\widehat{\boldsymbol{\beta}}_j^\flat) = \mathbf{0}$, we have

$$\mathrm{II}_0 = \| \mathbf{J}_0(\widehat{\boldsymbol{\beta}}_0^\flat - \boldsymbol{\beta}_0^*) - \{\widehat{Q}_0^\flat(\widehat{\boldsymbol{\beta}}_0^\flat) - \widehat{Q}_0^\flat(\boldsymbol{\beta}_0^*)\} \|_{\Sigma^{-1}}$$

and for $j = 1, \ldots, m$,

$$\begin{aligned}\mathrm{II}_j \le{}& \| \mathbf{J}_j(\widehat{\boldsymbol{\beta}}_j^\flat - \boldsymbol{\beta}_j^*) - \{\widehat{Q}_j^\flat(\widehat{\boldsymbol{\beta}}_j^\flat) - \widehat{Q}_j^\flat(\boldsymbol{\beta}_j^*)\} \|_{\Sigma^{-1}} \\ &+ \left\| \frac{1}{n} \sum_{i=1}^n \sum_{\ell=0}^{j-1} \int_{\tau_\ell}^{\tau_{\ell+1}} \mathrm{d}H(u) \Big[\mathbf{H}_\ell(\widehat{\boldsymbol{\beta}}_\ell - \boldsymbol{\beta}_\ell^*) - W_i \big\{ \bar{K}_h(y_i - \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_\ell^*) + \bar{K}_h(y_i - \boldsymbol{x}_i^{\mathrm{T}} \widehat{\boldsymbol{\beta}}_\ell) \boldsymbol{x}_i \big\} \Big] \right\|_{\Sigma^{-1}}.\end{aligned}$$

Putting together the pieces, and taking $r \asymp \sqrt{(p + \log n)/n}$, we conclude that conditioning on $\mathscr{E}_1 \cap \mathscr{E}_2$,

$$\begin{aligned}\mathrm{II}_j \le{}& \sup_{\boldsymbol{\beta} \in \boldsymbol{\beta}_j^* + \Theta(r)} \| \mathbf{J}_j(\boldsymbol{\beta} - \boldsymbol{\beta}_j^*) - \{\widehat{Q}_j^\flat(\boldsymbol{\beta}) - \widehat{Q}_j^\flat(\boldsymbol{\beta}_j^*)\} \|_{\Sigma^{-1}} \qquad \text{(C.40)} \\ &+ \sup_{\cap_{\ell=0}^{j-1} \{\boldsymbol{\beta}_\ell \in \boldsymbol{\beta}_\ell^* + \Theta(r)\}} \left\| \frac{1}{n} \sum_{i=1}^n \sum_{\ell=0}^{j-1} \int_{\tau_\ell}^{\tau_{\ell+1}} \mathrm{d}H(u) \Big[\mathbf{H}_\ell(\boldsymbol{\beta}_\ell - \boldsymbol{\beta}_\ell^*) - W_i \big\{ \bar{K}_h(y_i - \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_\ell^*) + \bar{K}_h(y_i - \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_\ell) \boldsymbol{x}_i \big\} \Big] \right\|_{\Sigma^{-1}}\end{aligned}$$

holds with $\mathbb{P}^*$-probability at least $1 - C_3 n^{-1}$.

Let $\mathscr{E}_3$–$\mathscr{E}_5$ be the events from Lemmas C.3.3–C.3.5, so that $\mathbb{P}(\mathscr{E}_3 \cap \mathscr{E}_4 \cap \mathscr{E}_5) \ge 1 - C_4 n^{-1}$. Applying Lemmas C.3.4 and C.3.5 to (C.40) yields that for any $j = 0, 1, \ldots, m$,

$$\begin{aligned}\mathrm{II}_j \lesssim{}& \left\{ m_4^{1/2} \frac{p + \log n}{n h^{1/2}} + h \sqrt{\frac{p + \log n}{n}} \right. \qquad \text{(C.41)} \\ &\left. + \zeta_p^2 \frac{(p + \log n)(p \log n)^{1/2}}{n^{3/2} h} + \max_{0 \le \ell \le j-1} \Gamma_\ell(r) \right\} \cdot \left\{ \log\left(\tfrac{1 - \tau_0}{1 - \tau_j}\right) \vee 1 \right\}\end{aligned}$$

holds with $\mathbb{P}^*$-probability at least $1 - n^{-2}$ conditioned on $\mathscr{E}_4 \cap \mathscr{E}_5$, where $\Gamma_\ell(r)$ is defined in (C.29). Note that $\log\left(\frac{1-\tau_0}{1-\tau_j}\right) \le \log\left(\frac{1-\tau_0}{1-\tau_m}\right)$ is bounded by a constant. For $\Gamma_\ell(r)$, it follows from

Lemma C.3.3 with $r \asymp \sqrt{(p+\log n)/n}$ that conditioned on $\mathscr{E}_3$,

$$\max_{0\le\ell\le m} \Gamma_\ell(r) \lesssim m_4^{1/2}\frac{p+\log n}{nh^{1/2}} + \zeta_p^2\frac{(p+\log n)(p\log n)^{1/2}}{n^{3/2}h} \tag{C.42}$$

holds with $\mathbb{P}^*$-probability at least $1-C_5n^{-1}$ provided $n \gtrsim \zeta_p^2\log n$.

Finally, define the event $\mathscr{F} = \mathscr{E}_1\cap\mathscr{E}_2\cap\mathscr{E}_3\cap\mathscr{E}_4\cap\mathscr{E}_5$, satisfying $\mathbb{P}(\mathscr{F}) \ge 1-Cn^{-1}$ for some constant $C>0$ independent of (n,p). Combining (C.39), (C.41) and (C.42) proves (C.37), as claimed.

STEP 2. The arguments from Stage Two in the proof of Theorem 4.3.2 can be similarly applied to bridge the gap between discrete and continuous uniform bounds. Thus the details are omitted. □

C.3.4 Proof of Theorem 4.3.6

Without loss of generality, we assume $||\boldsymbol{a}_n||_\Sigma = 1$ throughout the proof; otherwise, we first rescale the vectors $\boldsymbol{a}_n$ so that the same arguments apply.

Conditioned on the observed data $\mathbb{D}_n = \{(y_i,\Delta_i,\boldsymbol{x}_i)\}_{i=1}^n$, we have $\mathbb{E}^*\langle\boldsymbol{a}_n,\boldsymbol{U}_i^\flat(\tau)\rangle = 0$ for any $\tau\in[\tau_L,\tau_U]$. By the asymptotic (conditional) tightness established in Lemma C.3.6 and the central limit theorem, the limiting distribution of $\mathbb{G}_n^\flat(\cdot)$ given $\mathbb{D}_n$ is a zero-mean Gaussian process. Following the arguments in Appendix 1 of Lin, Wei and Ying [1993], it suffices to show that the conditional covariance function of $\mathbb{G}_n^\flat(\cdot)$ given $\mathbb{D}_n$ converges to $H(\cdot,\cdot)$ defined in (4.21), which is the limit of the (unconditional) covariance function of $\mathbb{G}_n(\cdot)$. To this end, for any $s,t\in[\tau_L,\tau_U]$, note that

$$\begin{aligned}
\mathrm{Cov}^*\left(\mathbb{G}_n^\flat(s),\mathbb{G}_n^\flat(t)\right) &= \mathbb{E}^*\{\mathbb{G}_n^\flat(s)\mathbb{G}_n^\flat(t)\}\\
&= \frac{1}{n}\mathbb{E}^*\left\{\sum_{i=1}^n\langle\boldsymbol{a}_n,e_i\boldsymbol{U}_i(s)\rangle\right\}\cdot\left\{\sum_{i=1}^n\langle\boldsymbol{a}_n,e_i\boldsymbol{U}_i(t)\rangle\right\}\\
&= \frac{1}{n}\sum_{i=1}^n\boldsymbol{a}_n^{\mathrm{T}}\boldsymbol{U}_i(s)\boldsymbol{U}_i(t)^{\mathrm{T}}\boldsymbol{a}_n \xrightarrow{\text{a.s.}} H(s,t),
\end{aligned}$$

where the almost sure convergence follows from the strong law of large numbers. This completes the proof. □

C.4 Proof of Theorem 4.4.1

For $j = 0, 1, \ldots,$ and $r, q > 0$, define

$$\psi_j(r,q) = \sup_{\boldsymbol{\beta} \in \boldsymbol{\beta}_j^* + \Theta(r) \cap \Lambda(q)} \left\| \frac{1}{n} \sum_{i=1}^n \left\{ \bar{K}_h(y_i - \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}) - \bar{K}_h(y_i - \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_j^*) \right\} \boldsymbol{x}_i \right\|_\infty, \tag{C.43}$$

where $\Lambda(q) := \{\boldsymbol{u} \in \mathbb{R}^p : \|\boldsymbol{u}\|_1 \leq q \|\boldsymbol{u}\|_\Sigma\}$ is a cone-like set.

C.4.1 Technical lemmas

Lemma C.4.1. *Let $j = 0, 1, \ldots, m$, and $t > 0$.*

(i) With probability at least $1 - e^{-t}$,

$$\|\widehat{Q}_0(\boldsymbol{\beta}_j^*) - Q_0(\boldsymbol{\beta}_j^*)\|_\infty \leq \bar{\tau}_0 \left\{ \sigma \sqrt{\frac{2t + 2\log(2p)}{n}} + \frac{t + \log(2p)}{3n} \right\},$$

where $\bar{\tau}_0 = \max(\tau_0, 1 - \tau_0)$ and $\sigma^2 = \max_{1 \leq k \leq p} \sigma_{kk}$.

(ii) With probability at least $1 - e^{-t}$,

$$\left\| \frac{1}{n} \sum_{i=1}^n (1 - \mathbb{E}) \bar{K}_h(y_i - \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_j^*) \boldsymbol{x}_i \right\|_\infty \leq \sigma \sqrt{\frac{2t + 2\log(2p)}{n}} + \frac{t + \log(2p)}{3n}.$$

Proof. It suffices to prove part (i) since the second inequality can be obtained from the same argument. Fix j, we have

$$\|\widehat{Q}_0(\boldsymbol{\beta}_j^*) - Q_0(\boldsymbol{\beta}_j^*)\|_\infty = \max_{1 \leq k \leq p} \left| \frac{1}{n} \sum_{i=1}^n (1 - \mathbb{E}) \underbrace{\left\{ \Delta_i \bar{K}_h(\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_j^* - y_i) - \tau_0 \right\}}_{=: \xi_{ij}} x_{ik} \right|,$$

where $\xi_{ij} = \Delta_i \bar{K}_h(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_j^* - y_i) - \tau_0$ is such that $|\xi_{ij}| \le \bar{\tau}_0 = \max(\tau_0, 1-\tau_0)$ and $\mathbb{E}(\xi_{ij}x_{ik})^2 \le \bar{\tau}_0^2\sigma_{kk}$. Applying Bernstein's inequality yields that, with probability at least $1-2e^{-z}$,

$$\left|\frac{1}{n}\sum_{i=1}^{n}(1-\mathbb{E})\xi_{ij}x_{ik}\right| \le \bar{\tau}_0\left(\sigma_{kk}^{1/2}\sqrt{\frac{2z}{n}} + \frac{z}{3n}\right) \quad \text{for any } 1 \le k \le p.$$

The claimed bound then follows by taking $z = t + \log(2p)$ and the union bound. □

Lemma C.4.2. *For any $t > 0$, we have that with probability at least $1 - e^{-t}$,*

$$\psi_j(r,q) \lesssim \left(\frac{q}{h}\sqrt{\frac{\log p}{n}} + \bar{f}^{1/2}\sqrt{\frac{t+\log p}{nh}}\right)\cdot r + \frac{t+\log p}{n}.$$

Proof. For any j fixed, and $k = 1, \ldots, p$, define after a change of variable $\boldsymbol{v} = \boldsymbol{\beta} - \boldsymbol{\beta}_j^*$ that

$$\psi_{j,k}(r,q) = \sup_{\boldsymbol{v}\in\Theta(r)\cap\Lambda(q)}\left|\frac{1}{n}\sum_{i=1}^{n}(1-\mathbb{E})\underbrace{\left\{\bar{K}_h(\varepsilon_{ij} - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{v}) - \bar{K}_h(\varepsilon_{ij})\right\}x_{ik}}_{=: g_{\boldsymbol{v}}(y_i,\boldsymbol{x}_i)}\right|,$$

where $\varepsilon_{ij} = y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_j^*$. Then $\psi_j(r,q) \le \max_{1\le k\le p}\psi_{j,k}(r,q)$. Note that $\sup_{\boldsymbol{v}}|g_{\boldsymbol{v}}(y_i,\boldsymbol{x}_i)| \le |x_{ik}| \le 1$. Let σ be any positive number such that $\sigma^2 \ge \sup_{\boldsymbol{v}\in\Theta(r)\cap\Lambda(q)}\mathbb{E}g_{\boldsymbol{v}}^2(y_i,\boldsymbol{x}_i)$. By Theorem 7.3 in Bousquet [2003]—an improved version of Talagrand's inequality, we obtain that for any $z > 0$,

$$\psi_{j,k}(r,q) \le \mathbb{E}\psi_{i,k}(r,q) + \sqrt{\left\{\sigma^2 + 2\mathbb{E}\psi_{i,k}(r,q)\right\}\frac{2z}{n}} + \frac{z}{3n} \tag{C.44}$$

holds with probability at least $1 - e^{-z}$. For the second moment $\mathbb{E}g_{\boldsymbol{v}}^2(y_i,\boldsymbol{x}_i)$, by a change of

variable and Minkowski's integral inequality we derive that

$$\begin{aligned}
\mathbb{E}g_{\boldsymbol{v}}^2(y_i,\boldsymbol{x}_i) &= \mathbb{E}\left[x_{ik}^2\int_{-\infty}^{\infty}\left\{\bar{K}_h(u-\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{v})-\bar{K}_h(u)\right\}^2 f_{y_i}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_j^*+u|\boldsymbol{x}_i)\mathrm{d}u\right] \\
&= h\mathbb{E}\left[x_{ik}^2\int_{-\infty}^{\infty}\left\{\bar{K}(v-\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{v}/h)-\bar{K}(v)\right\}^2 f_{y_i}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_j^*+vh|\boldsymbol{x}_i)\mathrm{d}v\right] \\
&\le \overline{f}h^{-1}\mathbb{E}\left[x_{ik}^2(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{v})^2\int_{-\infty}^{\infty}\left\{\int_0^1 K(v-w\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{v}/h)\mathrm{d}w\right\}^2\mathrm{d}v\right] \\
&\le \overline{f}h^{-1}\mathbb{E}\left(x_{ik}^2(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{v})^2\left[\int_0^1\left\{\int_{-\infty}^{\infty}K^2(v-w\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{v}/h)\mathrm{d}v\right\}^{1/2}\mathrm{d}w\right]^2\right) \\
&\le \kappa_u\overline{f}h^{-1}\mathbb{E}(x_{ik}\cdot\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{v})^2 \le \kappa_u\overline{f}h^{-1}r^2, \text{ valid for any } \boldsymbol{v}\in\Theta(r).
\end{aligned}$$

It remains to bound $\mathbb{E}\psi_{j,k}(r,q)$ in the concentration inequality (C.44). Note that $g_{\boldsymbol{v}}(y_i,\boldsymbol{x}_i)$ is (κ_u/h)-Lipschitz continuous in $\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{v}$, i.e., for any $\boldsymbol{v},\boldsymbol{v}'$, $|g_{\boldsymbol{v}}(y_i,\boldsymbol{x}_i)-g_{\boldsymbol{v}'}(y_i,\boldsymbol{x}_i)|\le(\kappa_u/h)|\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{v}-\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{v}'|$. Hence, it follows from Rademacher symmetrization and Talagrand's contraction principle that

$$\begin{aligned}
\mathbb{E}\psi_{j,k}(r,q) &\le 2\mathbb{E}\left\{\sup_{\boldsymbol{v}\in\Theta(r)\cap\Lambda(q)}\left|\frac{1}{n}\sum_{i=1}^n e_i\psi_{ij}(\boldsymbol{v})\right|\right\} \\
&\le 4\kappa_u\mathbb{E}\left\{\sup_{\boldsymbol{v}\in\Theta(r)\cap\Lambda(q)}\left|\frac{1}{nh}\sum_{i=1}^n e_i\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{v}\right|\right\} \le 4\kappa_u\frac{qr}{nh}\cdot\mathbb{E}\left\|\sum_{i=1}^n e_i\boldsymbol{x}_i\right\|_\infty,
\end{aligned}$$

where $e_1,\ldots,e_n$ are independent Rademacher variables. By Hoeffding's moment inequality,

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

where $\mathbb{E}_e$ denotes the expectation over $\{e_i\}_{i=1}^n$. Plugging this into the previous bound yields

$$\mathbb{E}\psi_{j,k}(r,q) \le 4\kappa_u\frac{qr}{h}\sqrt{\frac{2\log(2p)}{n}}.$$

Finally, the claimed result follows by taking $z = t + \log p$ in (C.44) and the union bound.

□

The following result extends the restricted strong property in Lemma C.2.4 to high dimensions. It follows from Proposition 4.2 in Tan, Wang and Zhou [2021] with slight modifications.

Lemma C.4.3. *Assume Conditions 4.3.1–4.3.3 hold, and let $h, r > 0$ satisfy $4\eta_{1/4} r \le h \le \underline{g}/(2l_1)$ with $\eta_{1/4}$ defined in* (C.2)*. Then, for any $0 \le j \le m$ and $t > 0$,*

$$\mathbb{P}\left\{ D(\boldsymbol{\beta}, \boldsymbol{\beta}_j^*) \ge \frac{1}{2}\underline{g}\kappa_l \cdot ||\boldsymbol{\beta} - \boldsymbol{\beta}_j^*||_\Sigma^2 \text{ for all } \boldsymbol{\beta} \in \boldsymbol{\beta}_j^* + \Theta(r) \cap \Lambda(q) \right\} \ge 1 - e^{-t}$$

provided that $n \gtrsim h(q/r)^2(t \vee \log p)$.

C.4.2 Proof of the theorem

Following the argument as in the proof of Theorem 4.3.1, it suffices to derive a uniform bound on the grid of τ-levels, $\tau_L = \tau_0 < \tau_1 < \cdots < \tau_m = \tau_U$. Again, we start with constructing intermediate points $\{\widetilde{\boldsymbol{\beta}}_j = (1-u_j)\boldsymbol{\beta}_j^* + u_j \widehat{\boldsymbol{\beta}}_j\}_{j=0,1,\ldots,m}$ that satisfy $\widetilde{\boldsymbol{\beta}}_j \in \boldsymbol{\beta}_j^* + \Theta(r^\diamondsuit)$, where $r^\diamondsuit = h/(4\eta_{1/4})$. For each $\widehat{\boldsymbol{\beta}}_j$, by the first-order optimality condition, there exists some subgradient $\widehat{\boldsymbol{g}}_j \in \partial\|\widehat{\boldsymbol{\beta}}_j\|_1$ such that $\widehat{Q}_j(\widehat{\boldsymbol{\beta}}_j) + \lambda_j \cdot \widehat{\boldsymbol{g}}_j = \mathbf{0}$ and $\langle \widehat{\boldsymbol{g}}_j, \widehat{\boldsymbol{\beta}}_j \rangle = \|\widehat{\boldsymbol{\beta}}_j\|_1$. Consequently, for each $j = 0, 1, \ldots, m$,

$$\begin{aligned} D(\widetilde{\boldsymbol{\beta}}_j, \boldsymbol{\beta}_j^*) &\le u_j D(\widehat{\boldsymbol{\beta}}_j, \boldsymbol{\beta}_j^*) = u_j \langle -\lambda_j \widehat{\boldsymbol{g}}_j - \widehat{Q}_j(\boldsymbol{\beta}_j^*), \widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^* \rangle \\ &\le \lambda_j \big(\|(\widetilde{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^*)_{\mathscr{S}_j}\|_1 - \|(\widetilde{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^*)_{\mathscr{S}_j^c}\|_1 \big) + \langle -\widehat{Q}_j(\boldsymbol{\beta}_j^*), \widetilde{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^* \rangle, \end{aligned} \tag{C.45}$$

where $\mathscr{S}_j = \mathrm{supp}(\boldsymbol{\beta}_j^*)$. Denote the cardinality of $\mathscr{S}_j$ by s_j, satisfying $s_j \le s$ for all j. Consider the decomposition

$$\widehat{Q}_j(\boldsymbol{\beta}_j^*) = \widehat{Q}_0(\boldsymbol{\beta}_j^*) - Q_0(\boldsymbol{\beta}_j^*) - \sum_{\ell=1}^{j-1} w_\ell(\widehat{\Delta}_\ell + \Delta_\ell) + Q_0(\boldsymbol{\beta}_j^*) - \sum_{\ell=0}^{j-1} w_\ell \mathbb{E}\{\bar{K}_h(y - \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_\ell^*)\boldsymbol{x}\},$$

where $\widehat{\Delta}_\ell$ are Δ_ℓ ($\ell = 0,1,\ldots,m-1$) are given in (C.12). Then, by Hölder's inequality,

$$
\begin{aligned}
&|\langle \widehat{Q}_j(\boldsymbol{\beta}_j^*), \widetilde{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^* \rangle| \qquad\qquad (C.46)\\
&\le \left\{ \|\widehat{Q}_0(\boldsymbol{\beta}_j^*) - Q_0(\boldsymbol{\beta}_j^*)\|_\infty + \sum_{\ell=0}^{j-1} w_\ell \big(\|\widehat{\Delta}_\ell\|_\infty + \|\Delta_\ell\|_\infty \big) \right\} \|\widetilde{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^*\|_1 \\
&\quad + \underbrace{\left\| Q_0(\boldsymbol{\beta}_j^*) - \sum_{\ell=0}^{j-1} w_\ell \mathbb{E}\{\bar{K}_h(y - \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_\ell^*)\boldsymbol{x}\} \right\|_{\Sigma^{-1}}}_{< (1+W_j)a \text{ by (C.14)}} \|\widetilde{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^*\|_\Sigma, \quad j = 1,\ldots,m,
\end{aligned}
$$

and $|\langle \widehat{Q}_0(\boldsymbol{\beta}_0^*), \widetilde{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}_0^* \rangle| \le \|\widehat{Q}_0(\boldsymbol{\beta}_0^*) - Q_0(\boldsymbol{\beta}_0^*)\|_\infty \|\widetilde{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}_0^*\|_1 + a\|\widehat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}_0^*\|_\Sigma$, where

$$
a = 0.5 l_1 \kappa_2 h^2 + \overline{f}\underline{f}^{-1}\delta^* \quad \text{and} \quad W_j = \sum_{\ell=0}^{j-1} w_\ell = \int_{\tau_L}^{\tau_j} \mathrm{d}H(u) = \log\big(\tfrac{1-\tau_L}{1-\tau_j}\big). \qquad (C.47)
$$

For some positive sequence $\{q_j\}_{j=0,1,\ldots,m}$ and curvature parameter $\kappa > 0$ to be determined, define the "good" events

$$
\begin{aligned}
\mathscr{G} &= \left\{ \|\widehat{Q}_0(\boldsymbol{\beta}_j^*) - Q_0(\boldsymbol{\beta}_j^*)\|_\infty \le \frac{\lambda_0}{2} \right\} \\
&\quad \bigcap \left\{ \|\widehat{Q}_0(\boldsymbol{\beta}_j^*) - Q_0(\boldsymbol{\beta}_j^*)\|_\infty + \sum_{\ell=0}^{j-1} w_\ell \|\Delta_\ell\|_\infty \le \frac{\lambda_j}{3},\ j = 1,\ldots,m \right\} \text{ and} \\
\mathscr{F} &= \bigcap_{j=0}^{m} \left\{ D(\boldsymbol{\beta}, \boldsymbol{\beta}_j^*) \ge \kappa \cdot \|\boldsymbol{\beta} - \boldsymbol{\beta}_j^*\|_\Sigma^2 \text{ for all } \boldsymbol{\beta} \in \boldsymbol{\beta}_j^* + \Theta(r^\diamondsuit) \cap \Lambda(q_j) \right\}.
\end{aligned}
$$

Conditioned on $\mathscr{F} \cap \mathscr{G}$, it follows from (C.45) that

$$
0 \le D(\widetilde{\boldsymbol{\beta}}_0, \boldsymbol{\beta}_0^*) < \frac{\lambda_0}{2}\big(3\|(\widetilde{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}_0^*)_{\mathscr{S}_0}\|_1 - \|(\widetilde{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}_0^*)_{\mathscr{S}_0^{\mathrm{c}}}\|_1\big) + a\|\widetilde{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}^*\|_\Sigma, \qquad (C.48)
$$

thus implying the cone-like constraint $\|(\widetilde{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}_0^*)_{\mathscr{S}_0^{\mathrm{c}}}\|_1 \le 3\|(\widetilde{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}_0^*)_{\mathscr{S}_0}\|_1 + (2a/\lambda_0)\|\widetilde{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}^*\|_\Sigma$. Taking $q_0 = 4(s_0/\gamma_l)^{1/2} + 2a/\lambda_0$, we see that $\widetilde{\boldsymbol{\beta}}_0$ falls into the cone-like set $\boldsymbol{\beta}_0^* + \Lambda(q_0)$, and so does $\widehat{\boldsymbol{\beta}}_0$. Hence, $D(\widetilde{\boldsymbol{\beta}}_0, \boldsymbol{\beta}_0^*) \ge \kappa \cdot \|\widetilde{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}_0^*\|_\Sigma^2$. Combining this with (C.48) yields, after

some algebra, that

$$\|\widetilde{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}_0^*\|_\Sigma < r_0 := \kappa^{-1}\big\{1.5(s/\gamma_l)^{1/2}\lambda_0 + a\big\}. \tag{C.49}$$

Provided $r_0 \le r^\diamondsuit$, $\widetilde{\boldsymbol{\beta}}_0$ lies in the interior of the local region $\boldsymbol{\beta}_0^* + \Theta(r^\diamondsuit)$. As before, we argue by contradiction that $\widehat{\boldsymbol{\beta}}_0$ coincides with $\widetilde{\boldsymbol{\beta}}_0$, thus implying $\widehat{\boldsymbol{\beta}}_0 \in \boldsymbol{\beta}_0^* + \Theta(r_0) \cap \Lambda(q_0)$.

For $(\widetilde{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_1)$, from (C.45) and (C.46) it follows that

$$\begin{aligned} 0 \le D(\widetilde{\boldsymbol{\beta}}_1, \boldsymbol{\beta}_1^*) < \lambda_1\big(\|(\widetilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*)_{\mathscr{S}_1}\|_1 - \|(\widetilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*)_{\mathscr{S}_1^c}\|_1\big) \\ + \big(\lambda_1/3 + w_0\|\widehat{\Delta}_0\|_\infty\big)\|\widetilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*\|_1 + (a + w_0 a)\|\widetilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*\|_\Sigma. \end{aligned}$$

We have already shown that $\widehat{\boldsymbol{\beta}}_0 \in \boldsymbol{\beta}_0^* + \Theta(r_0) \cap \Lambda(q_0)$ conditioning on $\mathscr{F} \cap \mathscr{G}$. Then $\|\widehat{\Delta}_0\|_\infty \le \psi_0(r_0, q_0)$, where $\psi_j(\cdot,\cdot)$ is defined in (C.43). Conditioned further on $\{w_0\psi_0(r_0,q_0) \le \lambda_1/3\}$, we have

$$0 \le D(\widetilde{\boldsymbol{\beta}}_1, \boldsymbol{\beta}_1^*) < \frac{\lambda_1}{3}\big(5\|(\widetilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*)_{\mathscr{S}_1}\|_1 - \|(\widetilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*)_{\mathscr{S}_1^c}\|_1\big) + a(1+w_0)\|\widetilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*\|_\Sigma,$$

which in turn implies $\widetilde{\boldsymbol{\beta}}_1 \in \boldsymbol{\beta}_1^* + \Lambda(q_1)$ with $q_1 := 6(s_1/\gamma_l)^{1/2} + 3(1+w_0)a/\lambda_1$. On the event $\mathscr{F}$, $D(\widetilde{\boldsymbol{\beta}}_1, \boldsymbol{\beta}_1^*) \ge \kappa \cdot \|\widetilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*\|_\Sigma^2$. Combining the upper and lower bounds yields

$$\|\widetilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*\|_\Sigma < r_1 := \kappa^{-1}\big\{\tfrac{5}{3}(s/\gamma_l)^{1/2}\lambda_1 + a + w_0 a\big\}. \tag{C.50}$$

Provided $r_1 \le r^\diamondsuit$, we reach the conclusion that $\widehat{\boldsymbol{\beta}}_1 \in \boldsymbol{\beta}_1^* + \Theta(r_1) \cap \Lambda(q_1)$.

We now recurse this argument, in particular controlling the error terms $\|\widehat{\Delta}_\ell\|_\infty$ sequentially,

so that at the j-th step ($1 \le j \le m$), $\widetilde{\boldsymbol{\beta}}_j$ satisfies

$$0 \le D(\widetilde{\boldsymbol{\beta}}_j, \boldsymbol{\beta}_j^*) < \lambda_j\big(\|(\widetilde{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^*)_{\mathscr{S}_j}\|_1 - \|(\widetilde{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^*)_{\mathscr{S}_j^c}\|_1\big) + \left(\frac{\lambda_j}{3} + \sum_{\ell=0}^{j-1} w_\ell \|\widehat{\Delta}_\ell\|_\infty\right)\|\widetilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*\|_1 + a\left(1 + \sum_{\ell=0}^{j-1} w_\ell\right)\|\widetilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*\|_\Sigma.$$

Conditioning on the event $\{\sum_{\ell=0}^{j-1} w_\ell \psi_\ell(r_\ell, q_\ell) \le \lambda_j/3\}$, we obtain the cone-like constraint $\widetilde{\boldsymbol{\beta}}_j \in \boldsymbol{\beta}_j^* + \Lambda(q_j)$ with $q_j := 6(s_j/\gamma_l)^{1/2} + 3(1+W_j)a/\lambda_j$, thus implying

$$\|\widetilde{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^*\|_\Sigma < r_j := \kappa^{-1}\big\{\tfrac{5}{3}(s/\gamma_l)^{1/2}\lambda_j + a + W_j a\big\}. \tag{C.51}$$

As long as $\max_{0\le j\le m} r_j \le r^\diamondsuit$, we have established the result $\widehat{\boldsymbol{\beta}}_j \in \boldsymbol{\beta}_j^* + \Theta(r_j) \cap \Lambda(q_j)$ ($j = 0, 1, \ldots, m$) as a deterministic claim, conditioned on the event

$$\mathscr{F} \cap \mathscr{G} \cap \bigcap_{j=1}^{m} \left\{ \sum_{\ell=0}^{j-1} w_\ell \psi_\ell(r_\ell, q_\ell) \le \frac{\lambda_j}{3} \right\}$$

with properly chosen $\kappa > 0$ and parameters $\lambda_0, \lambda_1, \ldots, \lambda_m$, where $q_0 = 4(s_0/\gamma_l)^{1/2} + 2a/\lambda_0$ and $q_j = 6(s_j/\gamma_l)^{1/2} + 3(1+W_j)a/\lambda_j$ for $j \ge 1$ with W_j given in (C.47).

Next we choose $\{\lambda_j\}_{j=0,1,\ldots,m}$ in a sequential manner so that the above good event occurs with high probability. Applying the two inequalities in Lemma C.4.1, both with $t = 2\log p$, implies that with probability at least $1 - 2(m+1)p^{-2}$,

$$\|\widehat{Q}_0(\boldsymbol{\beta}_j^*) - Q_0(\boldsymbol{\beta}_j^*)\|_\infty + \sum_{\ell=0}^{j-1} w_\ell \|\Delta_\ell\|_\infty \lesssim (1+W_j)\sigma\sqrt{\frac{\log p}{n}} \ \text{ for all } j = 0, 1, \ldots, m,$$

where $W_0 \equiv 0$. Throughout, assume the following upper bound constraint on the magnitude of h:

$$h^2 \lesssim (s/\lambda_l)^{1/2}\sigma\sqrt{\frac{\log p}{n}}.$$

Starting at $j = 0$, set $\lambda_0 \asymp \sigma\sqrt{\log(p)/n}$ so that $q_0 \lesssim (s/\gamma_l)^{1/2}$ and $r_0 = \kappa^{-1}\{1.5(s/\gamma_l)^{1/2}\lambda_0 + a\} \lesssim \kappa^{-1}(s/\gamma_l)^{1/2}\lambda_0$. With this choice of (λ_0, r_0, q_0), it follows from Lemma C.4.2 with $t = 2\log p$ that, with probability at least $1 - p^{-2}$,

$$\psi_0(r_0, q_0) \lesssim \frac{s\lambda_0}{\kappa\gamma_l h}\sqrt{\frac{\log p}{n}} + \frac{\log p}{n}.$$

Recall that $a \asymp h^2 + \delta^* \lesssim h^2 + n^{-1/2}$. We then choose $\lambda_1 \asymp (1 + W_1)\sigma\sqrt{\log(p)/n}$ so that $\lambda_1 \geq 3\max\left\{w_0\psi_0(r_0, q_0), \|\widehat{Q}_0(\boldsymbol{\beta}_1^*) - Q_0(\boldsymbol{\beta}_1^*)\|_\infty + w_0\|\Delta_\ell\|_\infty\right\}$ as long as $h \gtrsim (\kappa\gamma_l)^{-1}w_0 s\sqrt{\log(p)/n}$. Furthermore, it follows that $q_1 \lesssim (s/\gamma_l)^{1/2}$ and $r_1 \asymp \kappa^{-1}(s/\gamma_l)^{1/2}\lambda_1$.

At a general $j \geq 1$, assume we already have $\lambda_\ell \asymp (1 + W_\ell)\sigma\sqrt{\log(p)/n}$, $q_\ell \lesssim (s/\gamma_l)^{1/2}$ and $r_\ell \asymp \kappa^{-1}(s/\gamma_l)^{1/2}\lambda_\ell$ for $\ell = 0, 1, \ldots, j-1$. And with probability at least $1 - jp^{-2}$,

$$\psi_\ell(r_\ell, q_\ell) \lesssim \frac{s\lambda_\ell}{\kappa\gamma_l h}\sqrt{\frac{\log p}{n}} + \frac{\log p}{n}, \quad \ell = 0, 1, \ldots, j-1.$$

The accumulated error can thus be bounded by

$$\sum_{\ell=0}^{j-1} w_\ell\psi_\ell(r_\ell, q_\ell) \lesssim \frac{\sigma}{\kappa\gamma_l h}\frac{s\log p}{n}\sum_{\ell=0}^{j-1}(1 + W_\ell)w_\ell + W_j\frac{\log p}{n}.$$

Provided that $h \gtrsim (\kappa\gamma_l)^{-1}W_j s\sqrt{\log(p)/n}$,

$$(1 + W_j)\sigma\sqrt{\frac{\log p}{n}} \asymp \lambda_j \geq 3\max\left\{\sum_{\ell=0}^{j-1} w_\ell\psi_\ell(r_\ell, q_\ell), \|\widehat{Q}_0(\boldsymbol{\beta}_j^*) - Q_0(\boldsymbol{\beta}_j^*)\|_\infty + \sum_{\ell=0}^{j-1} w_\ell\|\Delta_\ell\|_\infty\right\}.$$

and therefore the event that involves λ_j is certified.

With the above choice of $\{r_j\}_{j=0,1,\ldots,m}$ and the lower bound constraint on the magnitude of the bandwidth—$h \gtrsim (\kappa\gamma_l)^{-1}W_m s\sqrt{\log(p)/n}$, we have

$$\kappa^{-1}(1 + W_j)\sigma(s/\gamma_l)^{1/2}\sqrt{\frac{\log p}{n}} \asymp r_j \leq r^\diamondsuit \asymp h \text{ for all } j = 0, 1, \ldots, m.$$

Finally, by the restricted strong convexity lemma—Lemma C.4.3 with $r = h/(4\eta_{1/4})$, $q \asymp (s/\gamma_l)^{1/2}$ and $t = 2\log p$—we take $\kappa = (\underline{g}\kappa_l)/2$ so that event $\mathscr{F}$ happens with probability at least $1-(m+1)p^{-2}$ provided that the "effective" sample size satisfies $nh \gtrsim s\log p$. □

C.5 Proof of Technical Lemmas

This section contains the proofs of all technical lemmas from Sections C.2 and C.3.

C.5.1 Proof of Lemma C.2.1

Fix $\boldsymbol{\beta}, \boldsymbol{\beta}' \in \mathbb{R}^p$, and define the function $f(\eta) = \langle Q(\boldsymbol{\beta}_\eta) - Q(\boldsymbol{\beta}'), \boldsymbol{\beta} - \boldsymbol{\beta}' \rangle$ for $\eta \in [0,1]$. Since $Q(\cdot)$ is differentiable with a positive semi-definite Jacobian, we have $f'(\eta) = \langle \nabla Q(\boldsymbol{\beta}_\eta)(\boldsymbol{\beta} - \boldsymbol{\beta}'), \boldsymbol{\beta} - \boldsymbol{\beta}' \rangle \geq 0$, and hence $f(\cdot)$ is non-decreasing. Consequently, for any $\eta \in [0,1]$,

$$\begin{aligned} D(\boldsymbol{\beta}_\eta, \boldsymbol{\beta}^*) &= \langle Q(\boldsymbol{\beta}_\eta) - Q(\boldsymbol{\beta}'), \boldsymbol{\beta}_\eta - \boldsymbol{\beta}' \rangle = \eta \langle Q(\boldsymbol{\beta}_\eta) - Q(\boldsymbol{\beta}'), \boldsymbol{\beta} - \boldsymbol{\beta}' \rangle \\ &= \eta f(\eta) \leq \eta f(1) = \eta \langle Q(\boldsymbol{\beta}) - Q(\boldsymbol{\beta}'), \boldsymbol{\beta} - \boldsymbol{\beta}' \rangle = \eta D(\boldsymbol{\beta}, \boldsymbol{\beta}'), \end{aligned}$$

as claimed. □

C.5.2 Proof of Lemma C.2.2

First, by the variational representation of $\|\cdot\|_2$,

$$\Delta := \left\| \frac{1}{n} \sum_{i=1}^n (\xi_i \boldsymbol{z}_i - \mathbb{E}\xi_i \boldsymbol{z}_i) \right\|_2 = \sup_{\boldsymbol{u} \in \mathbb{S}^{p-1}} \frac{1}{n} \sum_{i=1}^n (1-\mathbb{E}) f_{\boldsymbol{u}}(\xi_i, \boldsymbol{z}_i),$$

where $f_{\boldsymbol{u}}(\xi_i, \boldsymbol{z}_i) := \langle \boldsymbol{u}, \xi_i \boldsymbol{z}_i \rangle$ satisfies $|f_{\boldsymbol{u}}(\xi_i, \boldsymbol{z}_i)| \leq M\zeta_p$ and $\mathbb{E}\{f_{\boldsymbol{u}}^2(\xi_i, \boldsymbol{z}_i)\} = \mathbb{E}\{\xi_i^2 \langle \boldsymbol{u}, \boldsymbol{z}_i \rangle^2\} \leq \sigma^2$. Applying a refined Talagrand's inequality (see, e.g., Theorem 7.3 in Bousquet [2003]) yields that

with probability at least $1-e^{-t}$,

$$\Delta \le 2\mathbb{E}\Delta + \sigma\sqrt{\frac{2t}{n}} + \frac{4M\zeta_p t}{3n}.$$

It remains to bound $\mathbb{E}\Delta$. By the Cauchy-Schwarz inequality,

$$\begin{aligned}\mathbb{E}\Delta &\le \left(\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n(\xi_i \boldsymbol{z}_i - \mathbb{E}\xi_i\boldsymbol{z}_i)\right\|_2^2\right)^{1/2} \le \frac{1}{n}\left(\sum_{i=1}^n \mathbb{E}\|\xi_i\boldsymbol{z}_i - \mathbb{E}\xi_i\boldsymbol{z}_i\|_2^2\right)^{1/2} \\ &\le \frac{1}{n}\left\{\sum_{i=1}^n \mathbb{E}\big(\xi_i^2\|\boldsymbol{z}_i\|_2^2\big)\right\}^{1/2} \le \frac{\sigma}{n^{1/2}}\big(\mathbb{E}\|\boldsymbol{z}_i\|_2^2\big)^{1/2} = \sigma\sqrt{\frac{p}{n}}.\end{aligned}$$

Combining the above two displays gives

$$\left\|\frac{1}{n}\sum_{i=1}^n(\xi_i\boldsymbol{z}_i - \mathbb{E}\xi_i\boldsymbol{z}_i)\right\|_2 \le 2\sigma\sqrt{\frac{p}{n}} + \sigma\sqrt{\frac{2t}{n}} + M\zeta_p\frac{4t}{3n}$$

holds with probability at least $1-e^{-t}$. □

C.5.3 Proof of Lemma C.2.3

Proof of (i). For each $j = 0, 1, \ldots, m$, define random variables $\xi_{ij} = \Delta_i \bar{K}_h(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_j^* - y_i) - \tau_0$, so that the centered process can be written as $\Sigma^{-1/2}\{\widehat{Q}_0(\boldsymbol{\beta}_j^*) - \mathbb{E}\widehat{Q}_0(\boldsymbol{\beta}_j^*)\} = (1/n)\sum_{i=1}^n(\xi_{ij}\boldsymbol{z}_i - \mathbb{E}\xi_{ij}\boldsymbol{z}_i)$. Since $\Delta_i \in \{0,1\}$ and $0 \le \bar{K}(\cdot) \le 1$, we have $|\xi_{ij}| \le \bar{\tau}_0 = \max(\tau_0, 1-\tau_0)$. In particular, for $j=0$, it is shown in the proof of Lemma C.2 in He et al. [2022] that $\mathbb{E}(\xi_{i0}^2|\boldsymbol{x}_i) \le \tau_0(1-\tau_0) + (1+\tau_0)l_1\kappa_2 h^2$. For general $j \ge 1$, we can simply use the crude second moment bound $\mathbb{E}(\xi_{ij}^2|\boldsymbol{x}_i) \le \bar{\tau}_0^2$. The claimed bound of (i) then follows directly from Lemma C.2.2.

Proof of (ii). The bound follows trivially from Lemma C.2.2 and the facts that $|\bar{K}_h(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_j^* - y_i)| \le 1$ and $\mathbb{E}\{\bar{K}_h^2(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_j^* - y_i)|\boldsymbol{x}_i\} \le 1$.

Proof of (iii). The proof is based on a similar argument used in the proof of Lemma C.2.2. Fix j, set $\boldsymbol{v} = \boldsymbol{\beta}_{j+1}^* - \boldsymbol{\beta}_j^*$ so that $\|\boldsymbol{v}\|_\Sigma \le \underline{f}^{-1}\delta^*$. By the monotonicity of $u \mapsto \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*(u)$ and $\bar{K}_h(\cdot)$, we

have

$$\left|\int_{\tau_j}^{\tau_{j+1}} \{\bar{K}_h(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*(u)-y_i)-\bar{K}_h(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*(\tau_j)-y_i)\}\mathrm{d}H(u)\right|$$
$$\leq w_j\{\bar{K}_h(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*_{j+1}-y_i)-\bar{K}_h(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*_j-y_i)\} \leq \kappa_u w_j h^{-1}\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{v} \leq \kappa_u \underline{f}^{-1} w_j h^{-1}\zeta_p\delta^*, \tag{C.52}$$

implying the boundedness, where the last step follows from Condition 4.3.4 and (4.10). To control the (conditional) second moment, note that

$$\begin{aligned}
&\mathbb{E}\left[\left\{\int_{\tau_j}^{\tau_{j+1}}\left\{\bar{K}_h(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*(u)-y_i)-\bar{K}_h(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*_j-y_i)\right\}\mathrm{d}H(u)\right\}^2\middle|\boldsymbol{x}_i\right]\\
&\leq w_j^2\int_{-\infty}^{\infty}\left\{\bar{K}_h(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*_{j+1}-u)-\bar{K}_h(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*_j-u)\right\}^2 f_y(u|\boldsymbol{x})\mathrm{d}u\\
&= w_j^2 h\int_{-\infty}^{\infty}\left\{\bar{K}(v+\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{v}/h)-\bar{K}(v)\right\}^2 f_y(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_j-hv|\boldsymbol{x})\mathrm{d}v\\
&\leq \overline{f} w_j^2 h^{-1}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{v})^2\int_{-\infty}^{\infty}\left\{\int_0^1 K(v+w\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{v}/h)\mathrm{d}w\right\}^2\mathrm{d}v\\
&\overset{(*)}{\leq} \overline{f} w_j^2 h^{-1}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{v})^2\left(\int_0^1\left\{\int_{-\infty}^{\infty}K^2(v+w\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{v}/h)\mathrm{d}v\right\}^{1/2}\mathrm{d}w\right)^2\\
&\leq \kappa_u\overline{f} w_j^2 h^{-1}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{v})^2,
\end{aligned}$$

where Minkowski's integral inequality is applied in step $(*)$. Turning to the unconditional second moment, we have for any unit vector $\boldsymbol{u}$ that

$$\begin{aligned}
&\mathbb{E}\left\{\int_{\tau_j}^{\tau_{j+1}}\left\{\bar{K}_h(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*(u)-y_i)-\bar{K}_h(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*(\tau_j)-y_i)\right\}\mathrm{d}H(u)\langle\boldsymbol{u},\boldsymbol{z}_i\rangle\right\}^2\\
&\qquad\leq \kappa_u\overline{f} w_j^2 h^{-1}\left\{\mathbb{E}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{v})^4\right\}^{1/2}\left\{\mathbb{E}(\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{u})^4\right\}^{1/2} \leq \kappa_u\overline{f}\underline{f}^{-2} m_4 w_j^2 h^{-1}\delta^{*2},
\end{aligned} \tag{C.53}$$

where m_4 is given in (4.9). Combining (C.52) and (C.53) with Talagrand's inequality as in Lemma C.2.2 proves the claimed bound. □

C.5.4 Proof of Lemma C.2.4

Throughout the proof, for any fixed $j = 0, 1, \ldots, m$, we write $\boldsymbol{\beta}^* = \boldsymbol{\beta}_j^*$, $\widehat{Q}(\cdot) = \widehat{Q}_j(\cdot)$ and $Q(\cdot) = \mathbb{E}\widehat{Q}(\cdot)$ for simplicity. Recall the smoothed estimating functions defined in (4.3), (4.4), and the induced metric (symmetrized Bregman divergence)

$$D(\boldsymbol{\beta}, \boldsymbol{\beta}^*) = \frac{1}{n}\sum_{i=1}^{n} \Delta_i \big\{ \bar{K}_h(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} - y_i) - \bar{K}_h(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^* - y_i) \big\} \boldsymbol{x}_i^{\mathrm{T}}(\boldsymbol{\beta} - \boldsymbol{\beta}^*), \tag{C.54}$$

where $\bar{K}_h(\cdot) = \bar{K}(\cdot/h)$. Given $h, r > 0$, define the events $\mathscr{E}_i = \{|\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^* - y_i| \le h/2\} \cap \{|\boldsymbol{x}_i^{\mathrm{T}}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)| \le \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_\Sigma \cdot h/(2r)\}$ for $i = 1, \ldots, n$. For any $\boldsymbol{\beta} \in \boldsymbol{\beta}^* + \Theta(r)$, it is easy to see that $|y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}| \le h$ on $\mathscr{E}_i$, hence implying

$$D(\boldsymbol{\beta}, \boldsymbol{\beta}^*) \ge \frac{\kappa_l}{nh}\sum_{i=1}^{n} \Delta_i \{\boldsymbol{x}_i^{\mathrm{T}}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)\}^2 \mathbb{1}_{\mathscr{E}_i}. \tag{C.55}$$

It then suffices to bound the right-hand side of (C.55) from below uniformly over $\boldsymbol{\beta} \in \boldsymbol{\beta}^* + \Theta(r)$.

For $R > 0$, define the function $\varphi_R(u) = u^2 \mathbb{1}(|u| \le R/2) + \{u\,\mathrm{sign}(u) - R\}^2 \mathbb{1}(R/2 < |u| \le R)$, which is R-Lipschitz continuous and satisfies the following properties: $\varphi_{cR}(cu) = c^2\varphi_R(u)$ for any $c \ge 0$, $\varphi_0(u) = 0$, and

$$u^2 \mathbb{1}(|u| \le R/2) \le \varphi_R(u) \le u^2 \mathbb{1}(|u| \le R). \tag{C.56}$$

For $\boldsymbol{\beta} \in \boldsymbol{\beta}^* + \Theta(r)$, consider the change of variable $\boldsymbol{\delta} = \Sigma^{1/2}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)/\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_\Sigma$. Together, (C.55) and (C.56) imply

$$\frac{D(\boldsymbol{\beta}, \boldsymbol{\beta}^*)}{\kappa_l \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_\Sigma^2} \ge D_0(\boldsymbol{\delta}) := \frac{1}{nh}\sum_{i=1}^{n} \omega_i \cdot \varphi_{h/(2r)}(\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\delta}), \tag{C.57}$$

where $\omega_i := \mathbb{1}(|\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^* - y_i| \le h/2, \Delta_i = 1)$.

We first bound the expectation $\mathbb{E}\{D_0(\boldsymbol{\delta})\}$, and then control the concentration of $D_0(\boldsymbol{\delta})$

around $\mathbb{E}\{D_0(\boldsymbol{\delta})\}$. When $0 < h \leq 1$, Condition 4.3.3 ensures that

$$\underline{g} \cdot h \leq \mathbb{E}(\omega_i | \boldsymbol{x}_i) = \int_{\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^* - h/2}^{\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^* + h/2} g(u|\boldsymbol{x}_i)\,\mathrm{d}u \leq \overline{g} \cdot h \ \text{ almost surely.} \tag{C.58}$$

It then follows from (C.56) and (C.58) that

$$\begin{aligned}\mathbb{E}\big\{\omega_i \cdot \varphi_{h/(2r)}(\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\delta})\big\} &\geq \underline{g}h \cdot \mathbb{E}\varphi_{h/(2r)}(\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\delta}) \geq \underline{g}h \cdot \mathbb{E}\big\{(\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\delta})^2 \mathbb{1}\big(|\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\delta}| \leq h/(4r)\big)\big\} \\ &= \underline{g}h \cdot \big\{1 - \mathbb{E}(\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\delta})^2 \mathbb{1}\big(|\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\delta}| > h/(4r)\big)\big\},\end{aligned}$$

which further implies

$$\inf_{\boldsymbol{\delta} \in \mathbb{S}^{p-1}} \mathbb{E}\{D_0(\boldsymbol{\delta})\} \geq \underline{g} \cdot \left[1 - \sup_{\boldsymbol{u} \in \mathbb{S}^{p-1}} \mathbb{E}\big\{(\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\delta})^2 \mathbb{1}\big(|\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\delta}| > h/(4r)\big)\big\}\right].$$

By the definition of η_ξ in (C.2), as long as $0 < r \leq h/(4\eta_{1/4})$,

$$\inf_{\boldsymbol{\delta} \in \mathbb{S}^{p-1}} \mathbb{E}\{D_0(\boldsymbol{\delta})\} \geq \frac{3}{4}\underline{g}. \tag{C.59}$$

Turning to the random process $\{D_0(\boldsymbol{\delta}) - \mathbb{E}D_0(\boldsymbol{\delta}) : \boldsymbol{\delta} \in \mathbb{S}^{p-1}\}$, it suffices to bound

$$\Lambda = \sup_{\boldsymbol{\delta} \in \mathbb{S}^{p-1}} \big\{-D_0(\boldsymbol{\delta}) + \mathbb{E}D_0(\boldsymbol{\delta})\big\}. \tag{C.60}$$

By the fact that $0 \leq \varphi_R(u) \leq \min\{(R/2)^2, (R/2)|u|\}$ for all $u \in \mathbb{R}$, we have

$$0 \leq (\omega_i/h)\varphi_{h/(2r)}(\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\delta}) \leq \omega_i \min\big\{(4r)^{-2}h, (4r)^{-1}|\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\delta}|\big\}.$$

Combining this with (C.58) yields

$$\mathbb{E}\big\{(\omega_i/h)^2 \varphi_{h/(2r)}^2(\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\delta})\big\} \leq (4r)^{-2}\mathbb{E}\big\{\mathbb{E}(\omega_i|\boldsymbol{x}_i)(\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\delta})^2\big\} \leq (4r)^{-2}\bar{g}h.$$

With the above preparations, we apply a refined Talagrand's inequality—Theorem 7.3 in Bousquet [2003]—to obtain that, for any $t > 0$,

$$\begin{aligned}\Lambda &\le \mathbb{E}\Lambda + (\mathbb{E}\Lambda)^{1/2}\sqrt{\frac{ht}{4r^2n}} + \bar{g}^{1/2}\sqrt{\frac{ht}{8r^2n}} + \frac{h}{(4r)^2}\frac{t}{3n} \\ &\le \frac{5}{4}\mathbb{E}\Lambda + \bar{g}^{1/2}\sqrt{\frac{ht}{8r^2n}} + (1/4+1/48)\frac{ht}{r^2n} \end{aligned} \tag{C.61}$$

with probability at least $1-e^{-t}$. It remains to bound $\mathbb{E}\Lambda$. To this end, we define

$$\mathscr{E}(\boldsymbol{\delta}; \boldsymbol{z}_i, y_i) = \frac{\omega_i}{h}\varphi_{h/(2r)}(\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\delta}) = \frac{1}{h}\varphi_{\omega_i h/(2r)}(\omega_i \boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\delta}), \quad \boldsymbol{\delta} \in \mathbb{S}^{p-1}.$$

where the second equality follows from the property that $\varphi_{cR}(cu) = c^2\varphi_R(u)$ for any $c \ge 0$. By the Lipschitz continuity of $\varphi_R(\cdot)$, $\mathscr{E}(\boldsymbol{\delta}; \boldsymbol{z}_i, y_i)$ is $(2r)^{-1}$-Lipschitz continuous in $\omega_i \boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\delta}$, and $\mathscr{E}(\boldsymbol{\delta}; \boldsymbol{z}_i, y_i) = 0$ for any $\boldsymbol{\delta}$ such that $\omega_i \boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\delta} = 0$. Furthermore, define the subset $T \subseteq \mathbb{R}^n$ as

$$T = \left\{\boldsymbol{t} = (t_1, \ldots, t_n)^{\mathrm{T}} : t_i = \omega_i \boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\delta}, i = 1, \ldots, n, \boldsymbol{\delta} \in \mathbb{S}^{p-1}\right\},$$

and contractions $\phi_i : \mathbb{R} \to \mathbb{R}$ as $\phi_i(t) = (2r/h) \cdot \varphi_{\omega_i h/(2r)}(t)$. In fact, the Lipschitz continuity of $\varphi_R(\cdot)$ implies $|\phi(t) - \phi(s)| \le |t-s|$ for all $t, s \in \mathbb{R}$. Let $\varepsilon_1, \ldots, \varepsilon_n$ be independent Rademacher random variables, and denote by $\mathbb{E}_\varepsilon$ the expectation taken only with respect to ε_i's. Then, via a standard symmetrization and contraction argument (see, e.g. Lemma 6.3 and Theorem 4.12 in Ledoux and Talagrand [1991]), we have

$$\begin{aligned}\mathbb{E}_\varepsilon\Lambda &\le 2\mathbb{E}_\varepsilon\left\{\sup_{\boldsymbol{\delta}\in\mathbb{S}^{p-1}} \frac{1}{n}\sum_{i=1}^n \varepsilon_i \mathscr{E}(\boldsymbol{\delta}; \boldsymbol{z}_i, y_i)\right\} = \frac{1}{r}\mathbb{E}_\varepsilon\left\{\sup_{\boldsymbol{t}\in T} \frac{1}{n}\sum_{i=1}^n \varepsilon_i \phi_i(t_i)\right\} \\ &\le \frac{1}{r}\mathbb{E}_\varepsilon\left(\sup_{\boldsymbol{t}\in T} \frac{1}{n}\sum_{i=1}^n \varepsilon_i t_i\right) = \frac{1}{r}\mathbb{E}_\varepsilon\left\{\sup_{\boldsymbol{\delta}\in\mathbb{S}^{p-1}} \frac{1}{n}\sum_{i=1}^n \varepsilon_i \cdot \omega_i \boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\delta}\right\} \le \frac{1}{r}\mathbb{E}_\varepsilon\left\|\frac{1}{n}\sum_{i=1}^n \varepsilon_i \omega_i \boldsymbol{z}_i\right\|_2.\end{aligned}$$

Taking the expectation over $\{(\boldsymbol{z}_i, y_i)\}_{i=1}^n$ on both sides yields $\mathbb{E}\Lambda \le \bar{g}^{1/2}\sqrt{hp/(r^2n)}$. Substituting

this into (C.61), we obtain

$$\Lambda \leq \bar{g}^{1/2} r^{-1} \left(\frac{5}{4} \sqrt{\frac{hp}{r^2 n}} + \sqrt{\frac{ht}{8r^2 n}} \right) + (1/4 + 1/48) r^{-2} \frac{ht}{n} \tag{C.62}$$

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

Finally, combining (C.57), (C.59), (C.60) and (C.62) completes the proof. □

C.5.5 Proof of Lemma C.2.5

To begin with, define the centered process

$$S(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} (1 - \mathbb{E}) \bar{K}_h(y_i - \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}) \boldsymbol{z}_i, \ \ \boldsymbol{\beta} \in \mathbb{R}^p.$$

After a change of variable $\boldsymbol{v} = \Sigma^{1/2}(\boldsymbol{\beta} - \boldsymbol{\beta}_j^*)$, we have

$$\sup_{\boldsymbol{\beta} \in \boldsymbol{\beta}_j^* + \Theta(r)} ||S(\boldsymbol{\beta}) - S(\boldsymbol{\beta}_j^*)||_2 = \sup_{\boldsymbol{v} \in \mathbb{B}^p(r)} \| \underbrace{S(\boldsymbol{\beta}_j^* + \Sigma^{-1/2} \boldsymbol{v}) - S(\boldsymbol{\beta}_j^*)}_{=: \Psi(\boldsymbol{v})} \|_2.$$

Since the empirical process $\Psi(\boldsymbol{v})$ is continuous with respect to $\boldsymbol{v}$, we will apply the concentration bound from Theorem A.3 in Spokoiny [2013] to control the supremum $\sup_{\boldsymbol{v} \in \mathbb{B}^p(r)} \|\Psi(\boldsymbol{v})\|_2$.

First, note that the function $\Psi(\cdot) : \mathbb{R}^p \to \mathbb{R}^p$ satisfies $\Psi(\mathbf{0}) = \mathbf{0}$, $\mathbb{E}\{\Psi(\boldsymbol{v})\} = \mathbf{0}$

$$\nabla \Psi(\boldsymbol{v}) = \frac{1}{n} \sum_{i=1}^{n} \big\{ \phi_{i,\boldsymbol{v}} \boldsymbol{z}_i \boldsymbol{z}_i^{\mathrm{T}} - \mathbb{E}\big(\phi_{\boldsymbol{v}} \boldsymbol{z} \boldsymbol{z}^{\mathrm{T}}\big) \big\},$$

where $\phi_{i,\boldsymbol{v}} = K_h(\boldsymbol{z}_i^{\mathrm{T}} \boldsymbol{v} - \varepsilon_i)$, $\phi_{\boldsymbol{v}} = K_h(\boldsymbol{z}^{\mathrm{T}} \boldsymbol{v} - \varepsilon)$ and $\varepsilon_i = y_i - \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_j^*$. It is easy to see that $0 \leq \phi_{i,\boldsymbol{v}} \leq \kappa_u / h$ with $\kappa_u = \sup_{u \in \mathbb{R}} K(u)$. For any $\boldsymbol{g}, \boldsymbol{h} \in \mathbb{S}^{p-1}$ and $|\lambda| \leq \min\{nh/(\kappa_u \zeta_p^2), n/\bar{g}\}$, by

independence and the elementary inequality $e^u \le 1+u+u^2 e^{|u|}/2$, we obtain that

$$\begin{aligned}
&\mathbb{E}\exp\{\lambda \boldsymbol{g}^{\mathrm{T}}\nabla\Psi(\boldsymbol{v})\boldsymbol{h}\} \\
&\le \left[1+\frac{\lambda^2}{2n^2}e^{\frac{\overline{g}|\lambda|}{n}\mathbb{E}|\boldsymbol{z}^{\mathrm{T}}\boldsymbol{g}\boldsymbol{z}^{\mathrm{T}}\boldsymbol{h}|}\mathbb{E}\{\phi_{\boldsymbol{v}}\boldsymbol{z}^{\mathrm{T}}\boldsymbol{g}\boldsymbol{z}^{\mathrm{T}}\boldsymbol{h}-\mathbb{E}(\phi_{\boldsymbol{v}}\boldsymbol{z}^{\mathrm{T}}\boldsymbol{g}\boldsymbol{z}^{\mathrm{T}}\boldsymbol{h})\}^2 e^{\frac{\kappa_u|\lambda|}{nh}|\boldsymbol{z}^{\mathrm{T}}\boldsymbol{g}\boldsymbol{z}^{\mathrm{T}}\boldsymbol{h}|}\right]^n \\
&\overset{\text{(i)}}{\le} \left[1+\frac{\lambda^2}{2n^2}e^{\overline{g}|\lambda|/n}\mathbb{E}\{\phi_{\boldsymbol{v}}\boldsymbol{z}^{\mathrm{T}}\boldsymbol{g}\boldsymbol{z}^{\mathrm{T}}\boldsymbol{h}-\mathbb{E}(\phi_{\boldsymbol{v}}\boldsymbol{z}^{\mathrm{T}}\boldsymbol{g}\boldsymbol{z}^{\mathrm{T}}\boldsymbol{h})\}^2 e^{\frac{\kappa_u|\lambda|}{nh}|\boldsymbol{z}^{\mathrm{T}}\boldsymbol{g}\boldsymbol{z}^{\mathrm{T}}\boldsymbol{h}|}\right]^n \\
&\le \left[1+\frac{e\lambda^2}{2n^2}\mathbb{E}\{\phi_{\boldsymbol{v}}\boldsymbol{z}^{\mathrm{T}}\boldsymbol{g}\boldsymbol{z}^{\mathrm{T}}\boldsymbol{h}-\mathbb{E}(\phi_{\boldsymbol{v}}\boldsymbol{z}^{\mathrm{T}}\boldsymbol{g}\boldsymbol{z}^{\mathrm{T}}\boldsymbol{h})\}^2 e^{\kappa_u\zeta_p^2|\lambda|/(nh)}\right]^n \\
&\le \left\{1+\frac{(e\lambda)^2}{2n^2}\mathbb{E}\big(\phi_{\boldsymbol{v}}\boldsymbol{z}^{\mathrm{T}}\boldsymbol{g}\boldsymbol{z}^{\mathrm{T}}\boldsymbol{h}\big)^2\right\}^n, \qquad \text{(C.63)}
\end{aligned}$$

where inequality (i) follows from the bound $\mathbb{E}|\boldsymbol{z}^{\mathrm{T}}\boldsymbol{g}\boldsymbol{z}^{\mathrm{T}}\boldsymbol{h}| \le 1$. For $\phi_{\boldsymbol{v}} = K_h(\boldsymbol{z}^{\mathrm{T}}\boldsymbol{v}-\varepsilon)$, under Condition 4.3.3, its conditional second moment can be bounded by

$$\begin{aligned}
\mathbb{E}\big(\phi_{\boldsymbol{v}}^2|\boldsymbol{x}\big) &= \frac{1}{h^2}\int_{-\infty}^{\infty} K^2\left(\frac{\boldsymbol{z}^{\mathrm{T}}\boldsymbol{v}+\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_j^*-t}{h}\right) f_y(t|\boldsymbol{x})\mathrm{d}t \\
&= \frac{1}{h}\int_{-\infty}^{\infty} K^2(u) f_y(\boldsymbol{z}^{\mathrm{T}}\boldsymbol{v}+\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_j^*+hu|\boldsymbol{x})\mathrm{d}u \le \frac{\kappa_u\overline{f}}{h}. \qquad \text{(C.64)}
\end{aligned}$$

Substituting this into (C.63) yields

$$\mathbb{E}\exp\{\lambda \boldsymbol{g}^{\mathrm{T}}\nabla\Psi(\boldsymbol{v})\boldsymbol{h}\} \le \{1+\kappa_u\overline{f}e^2 m_4\lambda^2/(2n^2h)\}^n \le \exp\{\kappa_u\overline{f}e^2 m_4\lambda^2/(2nh)\}.$$

This verifies condition (A.4) in Spokoiny [2013]. Therefore, applying Theorem A.3 therein, we obtain that with probability at least $1-e^{-t}$,

$$\begin{aligned}
&\sup_{\boldsymbol{\beta}\in\boldsymbol{\beta}_j^*+\Theta(r)}\left\|\frac{1}{n}\sum_{i=1}^n(1-\mathbb{E})\{\bar{K}_h(y_i-\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})-\bar{K}_h(y_i-\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_j^*)\}\boldsymbol{z}_i\right\|_2 \\
&\qquad = \sup_{\boldsymbol{v}\in\mathbb{B}^p(r)}\|\Psi(\boldsymbol{v})\|_2 \lesssim (\kappa_u\overline{f}m_4)^{1/2}\sqrt{\frac{p+t}{nh}}\cdot r \qquad \text{(C.65)}
\end{aligned}$$

as long as $nh \gtrsim \zeta_p^2(p+t)^{1/2}$. This proves (C.6), and (C.7) can be obtained from the same argument.

Turning to the mean difference approximation, applying the mean value theorem for vector-valued functions implies

$$\begin{aligned}
&\mathbb{E}\{\bar{K}_h(y-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta})-\bar{K}_h(y-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_j^*)\}\boldsymbol{z} \\
&= -\int_0^1 \mathbb{E}\big\{K_h(y-\langle \boldsymbol{x}, \boldsymbol{\beta}_j^* + t(\boldsymbol{\beta}-\boldsymbol{\beta}_j^*)\rangle)\boldsymbol{z}\boldsymbol{x}^{\mathrm{T}}\big\}\mathrm{d}t\cdot(\boldsymbol{\beta}-\boldsymbol{\beta}_j^*).
\end{aligned}$$

With $\boldsymbol{v} = \Sigma^{1/2}(\boldsymbol{\beta}-\boldsymbol{\beta}_j^*)$, note that

$$\begin{aligned}
&\mathbb{E}\big\{K_h(y-\langle \boldsymbol{x}, \boldsymbol{\beta}_j^* + t(\boldsymbol{\beta}-\boldsymbol{\beta}_j^*)\rangle)|\boldsymbol{x}\big\} \\
&= \frac{1}{h}\int_{-\infty}^{\infty} K\left(\frac{u-t\boldsymbol{z}^{\mathrm{T}}\boldsymbol{v}}{h}\right) f_y(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_j^* + u|\boldsymbol{x})\mathrm{d}u = \int_{-\infty}^{\infty} K(v) f_y(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_j^* + t\boldsymbol{z}^{\mathrm{T}}\boldsymbol{v} + hv|\boldsymbol{x})\mathrm{d}v.
\end{aligned}$$

By the Lipschitz continuity of $f_y(\cdot|\boldsymbol{x})$, we have

$$\begin{aligned}
&\left\|\mathbb{E}\{\bar{K}_h(y-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta})-\bar{K}_h(y-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_j^*)\}\boldsymbol{z} + \mathbb{E}\big\{f_y(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_j^*|\boldsymbol{x})\boldsymbol{z}\boldsymbol{z}^{\mathrm{T}}\big\}\boldsymbol{v}\right\|_2 \\
&= \left\|\mathbb{E}\int_0^1\int_{-\infty}^{\infty} K(v)\big\{f_y(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_j^* + t\boldsymbol{z}^{\mathrm{T}}\boldsymbol{v} + hv|\boldsymbol{x}) - f_y(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_j^*|\boldsymbol{x})\big\}\boldsymbol{z}\boldsymbol{z}^{\mathrm{T}}\mathrm{d}v\mathrm{d}t\cdot\boldsymbol{v}\right\|_2 \\
&\le l_1 \sup_{\boldsymbol{u}\in\mathbb{S}^{p-1}} \mathbb{E}\int_0^1\int_{-\infty}^{\infty} K(v)\big(t|\boldsymbol{z}^{\mathrm{T}}\boldsymbol{v}| + h|v|\big)\mathrm{d}v\mathrm{d}t\cdot|\boldsymbol{z}^{\mathrm{T}}\boldsymbol{u}\boldsymbol{z}^{\mathrm{T}}\boldsymbol{v}| \le l_1\big(0.5m_3 r + \kappa_1 h\big)r,
\end{aligned}$$

as claimed. □

C.5.6 Proof of Lemma C.2.6

For any $\varepsilon \in (0, \tau_u - \tau_l)$, we divide the interval $[\tau_l, \tau_u]$ into $L := \lceil(\tau_u-\tau_l)/(2\varepsilon)\rceil + 1$ subintervals, centered at the points τ^k for $k \in [L]$, and each of length at most 2ε. For any

$\tau \in [\tau_l, \tau_u]$, there exists some k such that $|\tau - \tau^k| \le \varepsilon$, and hence

$$\begin{aligned}
&\left\| \frac{1}{n}\sum_{i=1}^{n}(1-\mathbb{E})\int_{\tau_l}^{\tau} \bar{K}_h(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*(u))\mathrm{d}H(u)\cdot \boldsymbol{x}_i \right\|_{\Sigma^{-1}} \\
\le &\left\| \frac{1}{n}\sum_{i=1}^{n}(1-\mathbb{E})\int_{\tau_l}^{\tau^k} \bar{K}_h(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*(u))\mathrm{d}H(u)\cdot \boldsymbol{x}_i \right\|_{\Sigma^{-1}} \\
&+ \left\| \frac{1}{n}\sum_{i=1}^{n}(1-\mathbb{E})\int_{\tau^k}^{\tau} \bar{K}_h(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*(u))\mathrm{d}H(u)\cdot \boldsymbol{x}_i \right\|_{\Sigma^{-1}} \\
\le &\left\| \frac{1}{n}\sum_{i=1}^{n}(1-\mathbb{E})\int_{\tau_l}^{\tau^k} \bar{K}_h(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*(u))\mathrm{d}H(u)\cdot \boldsymbol{z}_i \right\|_{2} + 2\zeta_p |H(\tau) - H(\tau^k)|.
\end{aligned}$$

For any given $k \in [L]$, applying Lemma C.2.2 yields that with probability at least $1 - e^{-v}$,

$$\left\| \frac{1}{n}\sum_{i=1}^{n}(1-\mathbb{E})\int_{\tau_l}^{\tau^k} \bar{K}_h(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*(u))\mathrm{d}H(u)\cdot \boldsymbol{z}_i \right\|_{2} \lesssim |H(\tau^k) - H(\tau_l)| \left(\sqrt{\frac{p+v}{n}} + \zeta_p \frac{v}{n} \right).$$

Recall that $H(u) = -\log(1-u)$, $u \in (0,1)$, we have $|H(u) - H(v)| \le |u-v|/(1-u\vee v)$. Finally, taking $\varepsilon = (\tau_u - \tau_l)/(2n)$, $v = \log L + t$ $(t > 0)$, and the union bound over $k = 1, \ldots, L$, we conclude that

$$\begin{aligned}
&\sup_{\tau \in [\tau_l, \tau_u]} \left\| \frac{1}{n}\sum_{i=1}^{n}(1-\mathbb{E})\int_{\tau_j}^{\tau} \bar{K}_h(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*(u))\mathrm{d}H(u)\cdot \boldsymbol{x}_i \right\|_{\Sigma^{-1}} \\
&\qquad \lesssim \frac{\tau_u - \tau_l}{1-\tau_u} \left(\sqrt{\frac{p + \log n + t}{n}} + \zeta_p \frac{\log n + t}{n} \right)
\end{aligned}$$

holds with probability at least $1 - e^{-t}$. This proves the claimed result. □

C.5.7 Proof of Lemma C.2.7

For $Q_0(\boldsymbol{\beta}) = \mathbb{E}\{\Delta \bar{K}_h(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta} - y) - \tau_0\}\boldsymbol{x}$, it follows from integration by parts and change of variables that

$$\begin{aligned}\mathbb{E}\{\Delta \bar{K}_h(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^* - y)|\boldsymbol{x}\} &= \int_{-\infty}^{\infty} \bar{K}\left(\frac{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^* - t}{h}\right) \mathrm{d}G(t|\boldsymbol{x}) = \frac{1}{h}\int_{-\infty}^{\infty} K\left(\frac{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^* - t}{h}\right) G(t|\boldsymbol{x})\mathrm{d}t \\ &= \int_{-\infty}^{\infty} K(u) G(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^* + hu|\boldsymbol{x})\mathrm{d}u \\ &= G(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*|\boldsymbol{x}) + \int_{-\infty}^{\infty} K(u) \int_{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*}^{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*+hu} \{g(t|\boldsymbol{x}) - g(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*|\boldsymbol{x})\}\mathrm{d}t\,\mathrm{d}u. \quad \text{(C.66)}\end{aligned}$$

On the other hand, using the martingale property gives

$$\mathbb{E}\left[\int_0^{\tau_j} \mathbb{1}\{y \geq \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*(u)\}\mathrm{d}H(u)\middle|\boldsymbol{x}\right] = \mathbb{E}\{N(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_j^*)|\boldsymbol{x}\} = \mathbb{P}(y \leq \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_j^*, \Delta = 1|\boldsymbol{x}) = G(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_j^*|\boldsymbol{x}).$$

Together, the last two displays and the Lipschitz continuity of $g(\cdot|\boldsymbol{x})$ imply

$$\left\|\mathbb{E}\left[\Delta \bar{K}_h(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_j^* - y) - \int_0^{\tau_j} \mathbb{1}\{y \geq \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*(u)\}\mathrm{d}H(u)\right]\boldsymbol{x}\right\|_{\Sigma^{-1}} \leq \frac{1}{2} l_1 \kappa_2 h^2. \quad \text{(C.67)}$$

Next, for $\ell = 0, 1, \ldots, j-1$,

$$\mathbb{E}\big\{\bar{K}_h(y - \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_\ell^*)|\boldsymbol{x}\big\} = 1 - F_y(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_\ell^*|\boldsymbol{x}) - \int_{-\infty}^{\infty} K(v) \int_{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_\ell^*}^{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_\ell^*+hv} \{f_y(t|\boldsymbol{x}) - f_y(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_\ell^*|\boldsymbol{x})\}\,\mathrm{d}t\,\mathrm{d}v.$$

This, combined with the Lipschitz continuity of $f(\cdot|\boldsymbol{x})$, implies

$$\left\|\mathbb{E}\left[\sum_{\ell=0}^{j-1} w_\ell \big\{\bar{K}_h(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_\ell^*) - \mathbb{1}(y_i \geq \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_\ell^*)\big\}\boldsymbol{x}_i\right]\right\|_{\Sigma^{-1}} \leq \frac{1}{2} l_1 \kappa_2 h^2 \sum_{\ell=0}^{j-1} w_\ell. \quad \text{(C.68)}$$

It remains to compare $\int_0^{\tau_j} \mathbb{1}\{y \geq \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*(u)\}\mathrm{d}H(u)$ and $\sum_{\ell=0}^{j-1} w_\ell \mathbb{1}(y_i \geq \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_\ell^*) + \tau_0$. By the global linear conditional quantile model assumption, the function $u \mapsto \mathbb{1}\{y \geq \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*(u)\}$ is non-increasing

in $u \in [\tau_L, \tau_U]$. Consequently,

$$\begin{aligned}
0 &\le \mathbb{E}\left(\sum_{\ell=0}^{j-1} \int_{\tau_\ell}^{\tau_{\ell+1}} \left[\mathbb{1}(y \ge \boldsymbol{x}^{\mathrm{T}} \boldsymbol{\beta}_\ell^*) - \mathbb{1}\{y \ge \boldsymbol{x}^{\mathrm{T}} \boldsymbol{\beta}^*(u)\} \right] \mathrm{d}H(u) \middle| \boldsymbol{x} \right) \\
&\le \mathbb{E}\left(\sum_{\ell=0}^{j-1} \int_{\tau_\ell}^{\tau_{\ell+1}} \left\{ \mathbb{1}(y \ge \boldsymbol{x}^{\mathrm{T}} \boldsymbol{\beta}_\ell^*) - \mathbb{1}(y \ge \boldsymbol{x}^{\mathrm{T}} \boldsymbol{\beta}_{\ell+1}^*) \right\} \mathrm{d}H(u) \middle| \boldsymbol{x} \right) \\
&= \sum_{\ell=0}^{j-1} \int_{\tau_\ell}^{\tau_{\ell+1}} \left\{ F_y(\boldsymbol{x}^{\mathrm{T}} \boldsymbol{\beta}_{\ell+1}^* | \boldsymbol{x}) - F_y(\boldsymbol{x}^{\mathrm{T}} \boldsymbol{\beta}_\ell^* | \boldsymbol{x}) \right\} \mathrm{d}H(u) \\
&\le \bar{f} \sum_{\ell=0}^{j-1} w_\ell \cdot \boldsymbol{x}^{\mathrm{T}} (\boldsymbol{\beta}_{\ell+1}^* - \boldsymbol{\beta}_\ell^*).
\end{aligned}$$

Recall from the last paragraph of Section 4.3.1 that $\|\boldsymbol{\beta}_{\ell+1}^* - \boldsymbol{\beta}_\ell^*\|_\Sigma \le \underline{f}^{-1} |\tau_{\ell+1} - \tau_\ell|$ for $\ell = 0, 1, \ldots, m-1$. Under the condition of no censoring below $\tau_0 = \tau_L$, we have $\mathbb{E} \int_0^{\tau_0} \mathbb{1}\{y \ge \boldsymbol{x}^{\mathrm{T}} \boldsymbol{\beta}^*(u)\} \mathrm{d}H(u) = \tau_0$. Putting together the pieces, we conclude that

$$\begin{aligned}
&\left\| \mathbb{E}\left[\int_0^{\tau_j} \mathbb{1}\{y \ge \boldsymbol{x}^{\mathrm{T}} \boldsymbol{\beta}^*(u)\} \mathrm{d}H(u) - \sum_{\ell=0}^{j-1} w_\ell \mathbb{1}(y \ge \boldsymbol{x}^{\mathrm{T}} \boldsymbol{\beta}_\ell^*) - \tau_0 \right] \boldsymbol{x} \right\|_{\Sigma^{-1}} \\
&\quad \le (\bar{f}/\underline{f}) \sum_{\ell=0}^{j-1} w_\ell (\tau_{\ell+1} - \tau_\ell).
\end{aligned}$$

Combining this with (C.67) and (C.68) proves (C.8). □

C.5.8 Proof of Lemma C.2.8

Fix $\delta > 0$, for any $\tau_1, \tau_2 \in [\tau_L, \tau_U]$ satisfying $|\tau_1 - \tau_2| < \delta$, define the centered random variables $V_i := \langle \boldsymbol{a}_n, \boldsymbol{U}_{0i}(\tau_1) - \boldsymbol{U}_{0i}(\tau_2) \rangle$ for $i = 1, \ldots, n$, where $\boldsymbol{U}_{0i}(\tau) = \boldsymbol{U}_i(\tau) - \mathbb{E}\boldsymbol{U}_i(\tau)$. Assume without loss of generality that $\tau_2 \ge \tau_1$. For some constant $L \ge 1$ to be determined, applying Rosenthal's inequality to $S := \sum_{i=1}^n V_i$ yields

$$\left(\mathbb{E}S^{2L}\right)^{1/(2L)} \le C_L \left\{ \left(\sum_{i=1}^n \mathbb{E}V_i^2 \right)^{1/2} + \left(\sum_{i=1}^n \mathbb{E}V_i^{2L} \right)^{1/(2L)} \right\},$$

where $C_L > 0$ is a constant depending only on L. We then bound the second and higher-order moments of V_i's. By Minkowski's integral inequality as in the proof of Lemma C.2.3,

$$\begin{aligned}
&\mathbb{E}\langle \boldsymbol{a}_n, \boldsymbol{U}_i(\tau_1) - \boldsymbol{U}_i(\tau_2)\rangle^2 \\
&= \mathbb{E}\bigg\{ -\int_{\tau_1}^{\tau_2} \bar{K}_h(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*(u))\mathrm{d}H(u) + \Delta_i \bar{K}_h(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*(\tau_2) - y_i) - \Delta_i \bar{K}_h(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*(\tau_1) - y_i) \bigg\}^2 \langle \boldsymbol{a}_n, \boldsymbol{x}_i\rangle^2 \\
&\lesssim (\overline{f} \vee \overline{g})\big\{(1-\tau_2)^{-2} + \underline{f}^{-2} m_4 h^{-1}\big\}(\tau_2 - \tau_1)^2.
\end{aligned}$$

Moreover, for any $q > 2$,

$$\begin{aligned}
&|\mathbb{E}\langle \boldsymbol{a}_n, \boldsymbol{U}_i(\tau_1) - \boldsymbol{U}_i(\tau_2)\rangle^q| \\
&\lesssim \zeta_p^{q/2-1}\mathbb{E}\langle \boldsymbol{a}_n, \boldsymbol{U}_i(\tau_1) - \boldsymbol{U}_i(\tau_2)\rangle^2 \lesssim (\overline{f} \vee \overline{g})\big\{(1-\tau_2)^{-2} + \underline{f}^{-2} m_4 h^{-1}\big\}\zeta_p^{q/2-1}(\tau_2 - \tau_1)^2.
\end{aligned}$$

Putting together the pieces, we obtain that for any $\tau_1 < \tau_2$ with $\tau_2 - \tau_1 \geq (\zeta_p/m_4)^{1/2}(h/n)^{1/2}$,

$$\begin{aligned}
\big(\mathbb{E}S^{2L}\big)^{1/(2L)} &\lesssim n^{1/2}(m_4/h)^{1/2}|\tau_2 - \tau_1| + n^{1/(2L)}\zeta_p^{(L-1)/(2L)}\big\{(m_4/h)^{1/2}|\tau_2 - \tau_1|\big\}^{1/L} \\
&\lesssim n^{1/2}(m_4/h)^{1/2}|\tau_2 - \tau_1|.
\end{aligned}$$

Note that $\mathbb{G}_n(\tau_1) - \mathbb{G}_n(\tau_2) = n^{-1/2}S$. Hence, taking $\psi(x) = x^{2L}$ in the above inequality leads to

$$||\mathbb{G}_n(\tau_1) - \mathbb{G}_n(\tau_2)||_\psi \lesssim (m_4/h)^{1/2}|\tau_2 - \tau_1|, \tag{C.69}$$

where $||\cdot||_\psi$ denotes the ψ-Orlicz norm; see Section 2.2 in van der Vaart and Wellner [1996].

The rest of the proof is based on a packing argument, and is inspired by the proof of Lemma A.3 in Chao, Volgushev and Cheng [2017]. Define the metric $d(\cdot,\cdot)$ as $d(s,t) = h^{-1/2}|s-t|$ for $s,t \in [\tau_L, \tau_U]$. Then, for any $\varepsilon > 0$, the packing number $\mathscr{P}([\tau_L, \tau_U], \varepsilon, d) \lesssim h^{1/2}\varepsilon^{-1}$. Let $\bar{\eta} = 2\sqrt{\zeta_p/(m_4 n)}$, so that $\lim_{n\to\infty}\bar{\eta} \to 0$, and (C.69) holds for all τ_1, τ_2 satisfying

$d(\tau_1,\tau_2) \geq \bar{\eta}/2$. Applying Lemma A.1 in the Appendix of Kley et al. [2016] gives

$$\begin{aligned}\sup_{|\tau_1-\tau_2|<\delta} \left|\mathbb{G}_n(\tau_1)-\mathbb{G}_n(\tau_2)\right| &= \sup_{d(\tau_1,\tau_2)<h^{-1/2}\delta} \left|\mathbb{G}_n(\tau_1)-\mathbb{G}_n(\tau_2)\right| \\ &\leq S_1 + \sup_{d(s,t)<\bar{\eta}, t\in\widetilde{\mathscr{T}}} \left|\mathbb{G}_n(s)-\mathbb{G}_n(t)\right|, \end{aligned} \tag{C.70}$$

where the set $\widetilde{\mathscr{T}}$ contains at most $\mathscr{P}([\tau_L,\tau_U],\bar{\eta},d) \lesssim h^{1/2}\bar{\eta}^{-1} \lesssim \sqrt{nh/\zeta_p}$ points, and S_1 is a random variable satisfying

$$\mathbb{P}(|S_1|>x) \lesssim \left\{ \int_{\bar{\eta}/2}^{\eta} \psi^{-1}(\mathscr{P}([\tau_L,\tau_U],\varepsilon,d))\mathrm{d}\varepsilon + (h^{-1/2}\delta + 2\bar{\eta})\psi^{-1}(\mathscr{P}^2([\tau_L,\tau_U],\eta,d)) \right\}^{2L} x^{-2L} \tag{C.71}$$

for any $\eta \geq \bar{\eta}$ and $x>0$. Note that

$$\int_{\bar{\eta}/2}^{\eta} \psi^{-1}(\mathscr{P}([\tau_L,\tau_U],\varepsilon,d))\mathrm{d}\varepsilon \lesssim h^{(4L)^{-1}} \int_{\bar{\eta}/2}^{\eta} \varepsilon^{-(2L)^{-1}}\mathrm{d}\varepsilon = h^{(4L)^{-1}} \cdot \frac{\eta^{1-(2L)^{-1}} - (\bar{\eta}/2)^{1-(2L)^{-1}}}{1-(2L)^{-1}},$$

and $\psi^{-1}(\mathscr{P}^2([\tau_L,\tau_U],\eta,d)) \lesssim h^{(2L)^{-1}}\eta^{-L^{-1}}$. Substituting these into (C.71) with $\eta = h^{-1/4}$ and $L=1$ implies

$$\begin{aligned}\mathbb{P}(|S_1|>x) &\lesssim \left\{ h^{(4L)^{-1}} \cdot \frac{\eta^{1-(2L)^{-1}} - (\bar{\eta}/2)^{1-(2L)^{-1}}}{1-(2L)^{-1}} + (h^{-1/2}\delta + 2\bar{\eta}) \cdot h^{(2L)^{-1}}\eta^{-L^{-1}} \right\}^{2L} \cdot x^{-2L} \\ &= \left\{ 2h^{1/8} - 2h^{1/4}(\bar{\eta}/2)^{1/2} + h^{1/4}\delta + 2\bar{\eta}h^{3/4} \right\}^2 \cdot x^{-2}. \end{aligned}$$

Since $\bar{\eta}\to 0$ and $h\to 0$ as $n\to\infty$, we have for any $x>0$ that

$$\lim_{\delta\to 0} \limsup_{n\to\infty} \mathbb{P}(|S_1|>x) = 0. \tag{C.72}$$

It remains to deal with the supremum on the right-hand side of (C.70). For any fixed

$t \in \widetilde{\mathscr{T}}$, and $s \in [\tau_L, \tau_U]$ satisfying $|s-t| < h^{1/2}\bar{\eta} = 2\sqrt{\zeta_p h/(m_4 n)}$,

$$
\begin{aligned}
&\sup_{|s-t|<h^{1/2}\bar{\eta}} \left|\mathbb{G}_n(s) - \mathbb{G}_n(t)\right| \\
\leq &\sup_{|s-t|<h^{1/2}\bar{\eta}} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \int_t^s \bar{K}_h(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*(u)) \mathrm{d}H(u) \cdot \boldsymbol{a}_n^{\mathrm{T}}\boldsymbol{x}_i \right| \\
&\quad + \sup_{|s-t|<h^{1/2}\bar{\eta}} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ \Delta_i \bar{K}_h(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*(t) - y_i) - \Delta_i \bar{K}_h(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*(s) - y_i) \right\} \boldsymbol{a}_n^{\mathrm{T}}\boldsymbol{x}_i \right| \\
:= &\, \mathrm{I} + \mathrm{II}.
\end{aligned} \tag{C.73}
$$

We bound the two terms on the right-hand side respectively. For the first one, applying the triangle inequality and Lemma C.2.6 to the centered random quantity yields

$$
\begin{aligned}
\mathrm{I} \leq &\sup_{|s-t|<h^{1/2}\bar{\eta}} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbb{E} \int_t^s \bar{K}_h(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*(u)) \mathrm{d}H(u) \cdot \boldsymbol{a}_n^{\mathrm{T}}\boldsymbol{x}_i \right| \\
&\quad + \sup_{|s-t|<h^{1/2}\bar{\eta}} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (1-\mathbb{E}) \int_t^s \bar{K}_h(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*(u)) \mathrm{d}H(u) \cdot \boldsymbol{a}_n^{\mathrm{T}}\boldsymbol{x}_i \right| \\
\lesssim &\, n^{-1/2} h^{1/2} \bar{\eta} \sum_{i=1}^{n} \mathbb{E}|\boldsymbol{a}_n^{\mathrm{T}}\boldsymbol{x}_i| + n^{1/2} h^{1/2} \bar{\eta} \left(\sqrt{\frac{p + \log n + v}{n}} + \zeta_p \frac{\log n + v}{n} \right) \\
\lesssim &\, (\zeta_p h)^{1/2} \left(\sqrt{\frac{p + \log n + v}{n}} + \zeta_p \frac{\log n + v}{n} \right)
\end{aligned} \tag{C.74}
$$

with probability at least $1 - e^{-v}$ for any $v > 0$. Turning to the second term, by the Lipschitz continuity of $\|\boldsymbol{\beta}^*(\cdot)\|_\Sigma$ as in (4.10), $|s-t| < h^{1/2}\bar{\eta}$ indicates $||\boldsymbol{\beta}^*(s) - \boldsymbol{\beta}^*(t)||_\Sigma \leq \underline{f}^{-1} h^{1/2} \bar{\eta} =$

$2\underline{f}^{-1}\sqrt{\zeta_p h/(m_4 n)}$. Denote $r := 2\underline{f}^{-1}\sqrt{\zeta_p h/(m_4 n)}$, we have

$$\begin{aligned}
\mathrm{II} &\le \sup_{|s-t|<h^{1/2}\bar{\eta}} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbb{E}\big\{ \Delta_i \bar{K}_h(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*(t) - y_i) - \Delta_i \bar{K}_h(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*(s) - y_i) \big\} \boldsymbol{a}_n^{\mathrm{T}} \boldsymbol{x}_i \right| \\
&\quad + \sup_{\boldsymbol{\beta} \in \boldsymbol{\beta}^*(t) + \Theta(r)} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (1-\mathbb{E}) \big\{ \Delta_i \bar{K}_h(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*(t) - y_i) - \Delta_i \bar{K}_h(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} - y_i) \big\} \boldsymbol{a}_n^{\mathrm{T}} \boldsymbol{x}_i \right| \\
&\lesssim (\zeta_p h)^{1/2} + \zeta_p^{1/2} \sqrt{\frac{p+\log n}{n}} \qquad \text{(C.75)}
\end{aligned}$$

with probability at least $1 - e^{-v}$ for any $v > 0$, where Lemma C.2.5–(ii) is applied in the last step. With the bandwidth $h \asymp \{(p+\log n)/n\}^{2/5}$ and under the sample size requirement $n \gtrsim \zeta_p^{5/2}(p+\log n)$, it follows from (C.73)–(C.75) with $v = 2\log n$ that, for any $t \in \widetilde{\mathscr{T}}$,

$$\sup_{s:|s-t|<h^{1/2}\bar{\eta}} \big| \mathbb{G}_n(s) - \mathbb{G}_n(t) \big| \lesssim \zeta_p^{1/2} \left(\frac{p+\log n}{n} \right)^{1/5}$$

holds with probability at least $1 - n^{-2}$. Since $|\widetilde{\mathscr{T}}| \lesssim \sqrt{nh/\zeta_p}$, taking the union bound over $t \in \widetilde{\mathscr{T}}$ renders

$$\sup_{d(s,t)<\bar{\eta}, t\in\widetilde{\mathscr{T}}} \big| \mathbb{G}_n(s) - \mathbb{G}_n(t) \big| = \sup_{|s-t|<h^{1/2}\bar{\eta}, t\in\widetilde{\mathscr{T}}} \big| \mathbb{G}_n(s) - \mathbb{G}_n(t) \big| \lesssim \zeta_p^{1/2} \left(\frac{p+\log n}{n} \right)^{1/5}$$

with probability at least $1 - \widetilde{c}/n$ for some constant $\widetilde{c} > 0$. Provided $\zeta_p^{5/2}(p+\log n) = o(n)$, this implies

$$\limsup_{n\to\infty} \mathbb{P}\left\{ \sup_{d(s,t)<\bar{\eta}, t\in\widetilde{\mathscr{T}}} \big| \mathbb{G}_n(s) - \mathbb{G}_n(t) \big| > x \right\} = 0, \quad \text{valid for any } x > 0. \qquad \text{(C.76)}$$

Finally, putting together (C.70), (C.72) and (C.76) completes the proof. □

C.5.9 Proof of Lemma C.3.1

For simplicity, we omit the subscript j as in the proof of Lemma C.2.4. Using arguments similar to those that lead (C.55) and (C.57), we obtain

$$\begin{aligned} D^\flat(\boldsymbol{\beta},\boldsymbol{\beta}^*) &= \frac{1}{n}\sum_{i=1}^n \Delta_i\big\{\bar{K}_h(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}-y_i)-\bar{K}_h(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*-y_i)\big\}(1+e_i)\boldsymbol{x}_i^{\mathrm{T}}(\boldsymbol{\beta}-\boldsymbol{\beta}^*) \\ &\geq \kappa_l\cdot\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_\Sigma^2\cdot\underbrace{\frac{1}{nh}\sum_{i=1}^n(1+e_i)\omega_i\cdot\varphi_{h/(2r)}(\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{v})}_{=:D_0^\flat(\boldsymbol{v})}, \end{aligned}$$

where $\boldsymbol{v}=\Sigma^{1/2}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)/\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_\Sigma\in\mathbb{S}^{p-1}$, $\omega_i=\Delta_i\mathbb{1}(|\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*-y_i|\leq h/2)\in\{0,1\}$, and $\varphi(\cdot)$ is as in (C.56). Recall the definition of $D_0(\boldsymbol{v})$ in (C.57), we have

$$\inf_{\boldsymbol{\beta}\in\boldsymbol{\beta}^*+\Theta(r)}\frac{D^\flat(\boldsymbol{\beta},\boldsymbol{\beta}^*)}{\kappa_l\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_\Sigma^2}\geq\inf_{\boldsymbol{v}\in\mathbb{S}^{p-1}}D_0(\boldsymbol{v})-\sup_{\boldsymbol{v}\in\mathbb{S}^{p-1}}\{D_0(\boldsymbol{v})-D_0^\flat(\boldsymbol{v})\}. \tag{C.77}$$

A lower bound for $\inf_{\boldsymbol{v}\in\mathbb{S}^{p-1}}D_0(\boldsymbol{v})$ can be derived from Lemma C.2.4. Let $\mathscr{E}_{\mathrm{rsc}}(t)$ be the event that the bounds in Lemma C.2.4 with $r=h/(4\eta_{1/4})$ hold uniformly over $j=0,1,\ldots,m$, so that $\mathbb{P}\{\mathscr{E}_{\mathrm{rsc}}(t)\}\geq 1-(m+1)e^{-t}$. It suffices to control the bootstrap error

$$\Gamma_n:=\sup_{\boldsymbol{v}\in\mathbb{S}^{p-1}}\{D_0(\boldsymbol{v})-D_0^\flat(\boldsymbol{v})\}=\sup_{\boldsymbol{\delta}\in\mathbb{S}^{p-1}}\frac{1}{nh}\sum_{i=1}^n e_i\omega_i\cdot\varphi_{h/(2r)}(\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{v}).$$

Since $\varphi_R(u)\leq(R/2)^2$, $\omega_i\in\{0,1\}$ and $e_i\in\{-1,1\}$, we have that $\mathbb{E}^*\{e_i\omega_i\cdot\varphi_{h/(2r)}(\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{v})\}^2\leq(h/4r)^4\omega_i$ and $|e_i\omega_i\cdot\varphi_{h/(2r)}(\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{v})|\leq(h/4r)^2$. Then, by Theorem 7.3 in Bousquet [2003],

$$\Gamma_n\leq 2\mathbb{E}^*(\Gamma_n)+\frac{h}{(4r)^2}\left\{\left(\frac{1}{n}\sum_{i=1}^n\omega_i\right)^{1/2}\sqrt{\frac{2t}{n}}+\frac{4t}{3n}\right\}$$

holds with probability at least $1-e^{-t}$. Furthermore, by the Lipschitz continuity of $u \to \varphi_R(u)$ and Ledoux-Talagrand contraction principle,

$$\begin{aligned}\mathbb{E}^*(\Gamma_n) &\le \frac{1}{2r}\mathbb{E}^*\left(\sup_{\boldsymbol{\delta}\in\mathbb{S}^{p-1}} \frac{1}{n}\sum_{i=1}^n e_i\omega_i \cdot \boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{v}\right) \\ &= \frac{1}{2r}\mathbb{E}^*\left\|\frac{1}{n}\sum_{i=1}^n e_i\omega_i\boldsymbol{z}_i\right\|_2 \le \frac{1}{2rn^{1/2}}\left(\frac{1}{n}\sum_{i=1}^n \omega_i\|\boldsymbol{z}_i\|_2^2\right)^{1/2}.\end{aligned}$$

Next we deal with the data-dependent quantities $(1/n)\sum_{i=1}^n \omega_i$ and $(1/n)\sum_{i=1}^n \omega_i\|\boldsymbol{z}_i\|_2^2$. Note that $\mathbb{E}(\omega_i|\boldsymbol{x}_i) \le \overline{g}h$ and thus $\mathbb{E}(\omega_i\|\boldsymbol{z}_i\|_2^2) \le \overline{g}ph$. Moreover, $\omega_i\|\boldsymbol{z}_i\|_2^2 \le \zeta_p^2$ (almost surely) and $\mathbb{E}(\omega_i^2\|\boldsymbol{z}_i\|_2^4) = \mathbb{E}(\omega_i\|\boldsymbol{z}_i\|_2^4) \le \overline{g}\zeta_p^2 ph$. By Bernstein's inequality, together

$$\frac{1}{n}\sum_{i=1}^n \omega_i \le \mathbb{E}\omega_i + \sqrt{\mathbb{E}\omega_i \cdot \frac{2t}{n}} + \frac{t}{3n} \le \left(\sqrt{\mathbb{E}\omega_i} + \sqrt{\frac{t}{2n}}\right)^2 \le \left(\sqrt{\overline{g}h} + \sqrt{\frac{t}{2n}}\right)^2 \tag{C.78}$$

and

$$\frac{1}{n}\sum_{i=1}^n \omega_i\|\boldsymbol{z}_i\|_2^2 \le \overline{g}ph + (\overline{g}ph)^{1/2}\zeta_p\sqrt{\frac{2t}{n}} + \zeta_p^2\frac{t}{3n} \le \frac{3}{2}\overline{g}ph + \zeta_p^2\frac{4t}{3n} \tag{C.79}$$

hold with probability at least $1-2e^{-t}$. Let $\mathscr{E}_{\mathrm{loc}}(t)$ be the event that (C.78) and (C.79) hold. Putting the above four bounds together, we conclude that conditioned on $\mathscr{E}_{\mathrm{loc}}(t)$,

$$\begin{aligned}\sup_{\boldsymbol{v}\in\mathbb{S}^{p-1}}\left\{D_0(\boldsymbol{v}) - D_0^\flat(\boldsymbol{v})\right\} \le &\left(\frac{3}{2}\overline{g}ph + \zeta_p^2\frac{4t}{3n}\right)^{1/2}\frac{1}{2rn^{1/2}} \\ &+ \frac{h}{(4r)^2}\left\{\left(\sqrt{\overline{g}h} + \sqrt{\frac{t}{2n}}\right)\sqrt{\frac{2t}{n}} + \frac{4t}{3n}\right\}\end{aligned} \tag{C.80}$$

holds with $\mathbb{P}^*$-probability greater than $1-e^{-t}$.

Define radius $r = h/(4\eta_{1/4})$ and event $\mathscr{E}_1(t) = \mathscr{E}_{\mathrm{rsc}}(t) \cap \mathscr{E}_{\mathrm{loc}}(t)$. Together, Lemma C.2.4,

(C.77) and (C.80) imply that, with $\mathbb{P}^*$-probability at least $1-e^{-t}$ conditioned on $\mathscr{E}_1(t)$,

$$D^{\flat}(\boldsymbol{\beta}, \boldsymbol{\beta}_j^*) \geq \frac{1}{2} \underline{g} \kappa_l \|\boldsymbol{\beta} - \boldsymbol{\beta}_j^*\|_{\Sigma}^2$$

holds uniformly over $\boldsymbol{\beta} \in \boldsymbol{\beta}_j^* + \Theta(r)$ as long as $nh \gtrsim \max(p, \zeta_p t^{1/2})$. Taking the union bound over $j = 0, 1, \ldots, m$ concludes the proof. □

C.5.10 Proof of Lemma C.3.2

We proceed the proof via a covering argument. Let $\mathscr{N}$ be an $(1/2)$-net of the unit sphere $\mathbb{S}^{p-1}$ with cardinality $|\mathscr{N}| \leq 5^p$ such that

$$\left\| \frac{1}{n} \sum_{i=1}^{n} (e_i \cdot \xi_i \boldsymbol{z}_i) \right\|_2 \leq 2 \max_{\boldsymbol{u} \in \mathscr{N}_\varepsilon} \frac{1}{n} \sum_{i=1}^{n} e_i \cdot \xi_i \langle \boldsymbol{u}, \boldsymbol{z}_i \rangle.$$

Since e_i's are independent Rademacher random variables and $|\xi_i| \leq M$, conditional on the data $\{\boldsymbol{x}_i, \xi_i\}_{i=1}^n$, it follows from Hoeffding's inequality (see, e.g., Theorem 2.8 in Boucheron, Lugosi and Massart [2013]) that, with $\mathbb{P}^*$-probability (over $\{e_i\}_{i=1}^n$) at least $1-e^{-u}$,

$$\frac{1}{n} \sum_{i=1}^{n} e_i \cdot \xi_i \langle \boldsymbol{u}, \boldsymbol{z}_i \rangle \leq M \lambda_{\max}^{1/2} \left(\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{z}_i \boldsymbol{z}_i^{\mathrm{T}} \right) \sqrt{\frac{2u}{n}}, \tag{C.81}$$

where $\lambda_{\max}(A)$ denotes the largest eigenvalue of a symmetric matrix A. For the normalized empirical design matrix $(1/n)\sum_{i=1}^n \boldsymbol{z}_i \boldsymbol{z}_i^{\mathrm{T}}$ satisfying $\mathbb{E}(\boldsymbol{z}_i \boldsymbol{z}_i^{\mathrm{T}}) = \mathbf{I}_p$ and $\|\boldsymbol{z}_i\|_2 \leq \zeta_p$ (almost surely), it follows from Theorem 5.41 in Vershynin [2012] that with probability at least $1-n^{-2}$,

$$\lambda_{\max}^{1/2} \left(\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{z}_i \boldsymbol{z}_i^{\mathrm{T}} \right) \leq 1 + C\zeta_p \sqrt{\frac{\log(2pn)}{n}}. \tag{C.82}$$

We denote by $\mathscr{E}_2$ the event that (C.82) holds.

Finally, taking a union bound over $\boldsymbol{u} \in \mathscr{N}$ and setting $u = 2(p + \log n)$ in (C.81), we

conclude that with $\mathbb{P}^*$-probability at least $1-n^{-2}$ conditioned on $\mathscr{E}_2$,

$$\left\|\frac{1}{n}\sum_{i=1}^{n}(e_i\cdot\xi_i\boldsymbol{z}_i)\right\|_2 \lesssim \left(1+\zeta_p\sqrt{\frac{\log n}{n}}\right)\sqrt{\frac{p+\log n}{n}} \lesssim \sqrt{\frac{p+\log n}{n}},$$

as long as the sample size satisfies $n \gtrsim \zeta_p^2 \log n$. This proves the claimed bound. □

C.5.11 Proof of Lemma C.3.3

To avoid unnecessary repetitions, we only provide details for bounding $\Gamma_j(r)$. After a change of variable $\boldsymbol{v} = \boldsymbol{\Sigma}^{1/2}(\boldsymbol{\beta}-\boldsymbol{\beta}_j^*) \in \mathbb{B}^p(r)$ for $\boldsymbol{\beta} \in \boldsymbol{\beta}_j^* + \Theta(r)$, we write

$$\Gamma_j^{\Delta}(r) = \sup_{\boldsymbol{v}\in\mathbb{B}^p(r)} \|G_j(\boldsymbol{v})\|_2 := \sup_{\boldsymbol{v}\in\mathbb{B}^p(r)} \left\|\frac{1}{n}\sum_{i=1}^{n} e_i\cdot\{\bar{K}_h(\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{v}-\varepsilon_{ij})-\bar{K}_h(-\varepsilon_{ij})\}\boldsymbol{z}_i\right\|_2, \tag{C.83}$$

where $\varepsilon_{ij} = y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_j^*$. The process $\{G_j(\cdot)\}$ satisfies $G_j(\mathbf{0})=\mathbf{0}$, $\mathbb{E}^*\{G_j(\boldsymbol{v})\}=\mathbf{0}$ and $\nabla G_j(\boldsymbol{v}) = (1/n)\sum_{i=1}^n e_i\phi_{ij,\boldsymbol{v}}\boldsymbol{z}_i\boldsymbol{z}_i^{\mathrm{T}}$, where $\phi_{ij,\boldsymbol{v}} = K_h(\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{v}-\varepsilon_{ij})$ and $K_h(x)=K(x/h)/h$. For any $\lambda\in\mathbb{R}$ and $\boldsymbol{u},\boldsymbol{w}\in\mathbb{S}^{p-1}$, we have

$$\begin{aligned}
&\mathbb{E}^*\exp\{\lambda n^{1/2}\boldsymbol{u}^{\mathrm{T}}\nabla G_j(\boldsymbol{v})\boldsymbol{w}\} = \prod_{i=1}^{n}\mathbb{E}^*\exp\{\lambda n^{-1/2}e_i\phi_{ij,\boldsymbol{v}}\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{u}\cdot\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{w}\}\\
&\le \prod_{i=1}^{n}\exp\left\{\frac{\lambda^2}{2n}\phi_{ij,\boldsymbol{v}}^2(\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{u}\cdot\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{w})^2\right\} = \exp\left\{\frac{\lambda^2}{2n}\sum_{i=1}^{n}\phi_{ij,\boldsymbol{v}}^2(\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{u}\cdot\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{w})^2\right\}.
\end{aligned}$$

Note that $\phi_{ij,\boldsymbol{v}} \le \kappa_u/h$, and by Hölder's inequality,

$$\frac{1}{n}\sum_{i=1}^{n}\phi_{ij,\boldsymbol{v}}^2(\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{u}\cdot\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{w})^2 \le \frac{\kappa_u}{h}\cdot\left\{\frac{1}{n}\sum_{i=1}^{n}\phi_{ij,\boldsymbol{v}}(\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{u})^4\right\}^{1/2}\left\{\frac{1}{n}\sum_{i=1}^{n}\phi_{ij,\boldsymbol{v}}(\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{w})^4\right\}^{1/2}.$$

Given $r>0$, define the supremum

$$\Psi(r) = \sup_{\boldsymbol{u},\boldsymbol{v}\in\mathbb{S}^{p-1}}\psi_{\boldsymbol{u},\boldsymbol{v}}(r) := \sup_{\boldsymbol{u},\boldsymbol{v}\in\mathbb{S}^{p-1}}\frac{1}{n}\sum_{i=1}^{n}K_h(r\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{v}-\varepsilon_{ij})(\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{u})^4.$$

Under this notation,

$$\sup_{\boldsymbol{v}\in\mathbb{B}^p(r)} \log \mathbb{E}^* \exp\{\lambda n^{1/2}\boldsymbol{u}^{\mathrm{T}}\nabla G_j(\boldsymbol{v})\boldsymbol{w}\} \le \frac{\kappa_u}{2h}\Psi(r)\lambda^2.$$

It then follows from a conditional version of Theorem A.3 in Spokoiny [2013] that

$$\sup_{\boldsymbol{v}\in\mathbb{B}^p(r)} \|G_j(\boldsymbol{v})\|_2 \lesssim \kappa_u^{1/2}\Psi(r)^{1/2}\cdot r\sqrt{\frac{p+\log n}{nh}} \tag{C.84}$$

holds with $\mathbb{P}^*$-probability at least $1-n^{-2}$.

It remains to control the data-dependent quantity $\Psi(r) = \sup_{\boldsymbol{u},\boldsymbol{v}\in\mathbb{S}^{p-1}} \psi_{\boldsymbol{u},\boldsymbol{v}}(r)$. For any $\varepsilon_1,\varepsilon_2\in(0,1)$, let $\{\boldsymbol{u}_1,\ldots,\boldsymbol{u}_{d_1}\}$ and $\{\boldsymbol{v}_1,\ldots,\boldsymbol{v}_{d_2}\}$ be the ε_1- and ε_2-nets of $\mathbb{S}^{p-1}$ with cardinalities $d_1\le(1+2/\varepsilon_1)^p$ and $d_2\le(1+2/\varepsilon_2)^p$. Given $\boldsymbol{u},\boldsymbol{v}\in\mathbb{S}^{p-1}$, there exist some $1\le\ell\le d_1$ and $1\le k\le d_2$ such that $\|\boldsymbol{u}-\boldsymbol{u}_\ell\|_2\le\varepsilon_1$ and $\|\boldsymbol{v}-\boldsymbol{v}_k\|_2\le\varepsilon_2$. Consider the decomposition

$$\psi_{\boldsymbol{u},\boldsymbol{v}}(r) = \psi_{\boldsymbol{u},\boldsymbol{v}}(r) - \psi_{\boldsymbol{u},\boldsymbol{v}_k}(r) + \psi_{\boldsymbol{u},\boldsymbol{v}_k}(r). \tag{C.85}$$

For $\psi_{\boldsymbol{u},\boldsymbol{v}}(r)-\psi_{\boldsymbol{u},\boldsymbol{v}_k}(r)$, the Lipschitz continuity of $K(\cdot)$ ensures that

$$|\psi_{\boldsymbol{u},\boldsymbol{v}}(r)-\psi_{\boldsymbol{u},\boldsymbol{v}_k}(r)| \le \frac{l_K r}{nh^2}\sum_{i=1}^n |\boldsymbol{z}_i^{\mathrm{T}}(\boldsymbol{v}-\boldsymbol{v}_k)|(\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{u})^4 \le \frac{l_K}{h^2}\zeta_p^3\lambda_{\max}\left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{z}_i\boldsymbol{z}_i^{\mathrm{T}}\right)\cdot r\varepsilon_2. \tag{C.86}$$

For $\psi_{\boldsymbol{u},\boldsymbol{v}_k}(r)$, using the triangle inequality gives

$$\begin{aligned}
\psi_{\boldsymbol{u},\boldsymbol{v}_k}(r)^{1/4} &= \left\{ \frac{1}{n}\sum_{i=1}^{n} K_h(r\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{v}_k - \varepsilon_{ij})\langle \boldsymbol{z}_i, \boldsymbol{u}_\ell + \boldsymbol{u} - \boldsymbol{u}_\ell\rangle^4 \right\}^{1/4} \\
&\le \left\{ \frac{1}{n}\sum_{i=1}^{n} K_h(r\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{v}_k - \varepsilon_{ij})(\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{u}_\ell)^4 \right\}^{1/4} \\
&\quad + \left\{ \frac{1}{n}\sum_{i=1}^{n} K_h(r\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{v}_k - \varepsilon_{ij})\langle \boldsymbol{z}_i, \boldsymbol{u} - \boldsymbol{u}_\ell\rangle^4 \right\}^{1/4} \\
&\le \psi_{\boldsymbol{u}_\ell,\boldsymbol{v}_k}(r) + \varepsilon_1 \cdot \sup_{\boldsymbol{u}\in\mathbb{S}^{p-1}} \psi_{\boldsymbol{u},\boldsymbol{v}_k}(r)^{1/4}.
\end{aligned}$$

Taking the supremum over $\boldsymbol{u} \in \mathbb{S}^{p-1}$ and then the maximum over ℓ yields

$$\sup_{\boldsymbol{u}\in\mathbb{S}^{p-1}} \psi_{\boldsymbol{u},\boldsymbol{v}_k}(r) \le (1-\varepsilon_1)^{-4} \max_{1\le\ell\le d_1} \psi_{\boldsymbol{u}_\ell,\boldsymbol{v}_k}(r). \tag{C.87}$$

The problem then boils down to controlling the maximum $\max_{(\ell,k)\in[d_1]\times[d_2]} \psi_{\boldsymbol{u}_\ell,\boldsymbol{v}_k}(r)$. Note that $\mathbb{E}K_h^2(r\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{v}_k - \varepsilon_{ij})(\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{u}_\ell)^8 \le \overline{f}\kappa_u m_4 \zeta_p^4/h$ and $K_h(r\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{v}_k - \varepsilon_{ij})(\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{u}_\ell)^4 \le \kappa_u\zeta_p^4/h$. By Bernstein's inequality, we have that with probability at least $1-e^{-u}$,

$$\begin{aligned}
\psi_{\boldsymbol{u}_\ell,\boldsymbol{v}_k}(r) &\le \mathbb{E}\psi_{\boldsymbol{u}_\ell,\boldsymbol{v}_k}(r) + (\overline{f}\kappa_u m_4)^{1/2}\zeta_p^2\sqrt{\frac{2u}{nh}} + \kappa_u\zeta_p^4\frac{u}{3nh} \\
&\le \overline{f}m_4 + (\overline{f}\kappa_u m_4)^{1/2}\zeta_p^2\sqrt{\frac{2u}{nh}} + \kappa_u\zeta_p^4\frac{u}{3nh} \\
&\le \frac{3}{2}\overline{f}m_4 + \kappa_u\zeta_p^4\frac{4u}{3nh}.
\end{aligned}$$

Taking $\varepsilon_1 = 1-2^{-1/4}$, $\varepsilon_2 = n^{-2}$ and $u = p\log(1+2/\varepsilon_1)(1+2/\varepsilon_2)+\log(n)$ in the above bounds, we conclude from (C.82) and (C.85)–(C.87) that with probability at least $1-n^{-1}$,

$$\Psi(r) \lesssim m_4 + \zeta_p^4\frac{p\log n}{nh} + \zeta_p^3\frac{r}{(nh)^2}$$

as long as $n \gtrsim \zeta_p^2\log n$. Substituting this bound into (C.84) completes the proof for a particular

$j \in \{0, \ldots, m\}$. The claimed uniform bound follows from a union bound over the grid points. $\square$

C.5.12 Proof of Lemma C.3.4

For $j = 0, \ldots, m$, define the random process

$$\Lambda_j^\flat(\boldsymbol{\beta}) = \widehat{Q}_j^\flat(\boldsymbol{\beta}) - \widehat{Q}_j^\flat(\boldsymbol{\beta}_j^*) - \mathbf{J}_j(\boldsymbol{\beta} - \boldsymbol{\beta}_j^*), \;\; \boldsymbol{\beta} \in \mathbb{R}^p.$$

The goal is to bound the local fluctuation $\sup_{\boldsymbol{\beta} \in \boldsymbol{\beta}_j^* + \Theta(r)} \|\Lambda_j^\flat(\boldsymbol{\beta})\|_2$ for a prespecified $r > 0$. Since $\mathbb{E}(W_i) = 1$, we have $\mathbb{E}^*\{\widehat{Q}_j^\flat(\boldsymbol{\beta})\} = \widehat{Q}_j(\boldsymbol{\beta})$, and $\mathbb{E}^*\{\Lambda_j^\flat(\boldsymbol{\beta})\} = \Lambda_j(\boldsymbol{\beta})$, where

$$\Lambda_j(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^n \Delta_i \big\{\bar{K}_h(\boldsymbol{x}_i^\mathrm{T}\boldsymbol{\beta} - y_i) - \bar{K}_h(\boldsymbol{x}_i^\mathrm{T}\boldsymbol{\beta}_j^* - y_i)\big\}\boldsymbol{x}_i - \mathbf{J}_j(\boldsymbol{\beta} - \boldsymbol{\beta}_j^*).$$

Consequently, by the triangle inequality,

$$\sup_{\boldsymbol{\beta} \in \boldsymbol{\beta}_j^* + \Theta(r)} \|\Lambda_j^\flat(\boldsymbol{\beta})\|_{\Sigma^{-1}} \le \underbrace{\sup_{\boldsymbol{\beta} \in \boldsymbol{\beta}_j^* + \Theta(r)} \|\Lambda_j^\flat(\boldsymbol{\beta}) - \Lambda_j(\boldsymbol{\beta})\|_{\Sigma^{-1}}}_{= \Gamma_j^\Delta(r) \text{ by (C.30)}} + \sup_{\boldsymbol{\beta} \in \boldsymbol{\beta}_j^* + \Theta(r)} \|\Lambda_j(\boldsymbol{\beta})\|_{\Sigma^{-1}}. \tag{C.88}$$

For the first term on the right-hand side of (C.88), Lemma C.3.3 guarantees the existence of an event $\mathscr{E}_3$ with $\mathbb{P}(\mathscr{E}_3) \ge 1 - 3n^{-1}$ such that, conditioned on $\mathscr{E}_3$,

$$\sup_{\boldsymbol{\beta} \in \boldsymbol{\beta}_j^* + \Theta(r)} \|\Lambda_j^\flat(\boldsymbol{\beta}) - \Lambda_j(\boldsymbol{\beta})\|_2 \lesssim \left(m_4^{1/2} + \zeta_p^2\sqrt{\frac{p\log n}{nh}}\right) \cdot r\sqrt{\frac{p + \log n}{nh}}, \tag{C.89}$$

holds with $\mathbb{P}^*$-probability at least $1 - n^{-2}$, provided $n \gtrsim \zeta_p^2 \log n$. Moreover, let $\mathscr{E}_3'$ be the event that (C.7) holds for all $j = 0, \ldots, m$ with $t = 2\log n$, so that $\mathbb{P}(\mathscr{E}_3') \ge 1 - (m+1)n^{-2}$. Conditioning on $\mathscr{E}_3'$,

$$\sup_{\boldsymbol{\beta} \in \boldsymbol{\beta}_j^* + \Theta(r)} \|\Lambda_j(\boldsymbol{\beta})\|_{\Sigma^{-1}} \lesssim \left(m_4^{1/2}\sqrt{\frac{p + \log n}{nh}} + m_3 r + h\right) \cdot r, \tag{C.90}$$

provided that $nh \gtrsim \zeta_p^2(p+t)$.

Finally, taking $\mathscr{E}_4 = \mathscr{E}_3 \cap \mathscr{E}_3'$ so that $\mathbb{P}(\mathscr{E}_4) \geq 1-(m+3n+1)n^{-2}$. Together, (C.88)–(C.90) yield that with $\mathbb{P}^*$-probability at least $1-n^{-2}$ conditioned on $\mathscr{E}_4$,

$$\sup_{\boldsymbol{\beta}\in\boldsymbol{\beta}_j^*+\Theta(r)} \|\Lambda_j^\flat(\boldsymbol{\beta})\|_2 \lesssim \left\{ m_4^{1/2}\sqrt{\frac{p+\log n}{nh}} + m_3 r + h + \zeta_p^2 \frac{(p\log n)^{1/2}(p+\log n)^{1/2}}{nh} \right\} \cdot r,$$

completing the proof. □

C.5.13 Proof of Lemma C.3.5

For $j = 1,\ldots,m$, define the random processes

$$\begin{aligned}
&R_j^\flat(\boldsymbol{\beta}_0,\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_{j-1}) \\
&= \frac{1}{n}\sum_{i=1}^n \sum_{\ell=0}^{j-1} \int_{\tau_\ell}^{\tau_{\ell+1}} \mathrm{d}H(u) \Big[W_i\big\{\bar{K}_h(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_\ell^*) - \bar{K}_h(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_\ell)\big\}\boldsymbol{z}_i - \Sigma^{-1/2}\mathbf{H}_\ell(\boldsymbol{\beta}_\ell - \boldsymbol{\beta}_\ell^*)\Big]
\end{aligned}$$

and

$$\begin{aligned}
&R_j(\boldsymbol{\beta}_0,\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_{j-1}) \\
&= \frac{1}{n}\sum_{i=1}^n \sum_{\ell=0}^{j-1} \int_{\tau_\ell}^{\tau_{\ell+1}} \mathrm{d}H(u) \Big[\big\{\bar{K}_h(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_\ell^*) - \bar{K}_h(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_\ell)\big\}\boldsymbol{z}_i - \Sigma^{-1/2}\mathbf{H}_\ell(\boldsymbol{\beta}_\ell - \boldsymbol{\beta}_\ell^*)\Big],
\end{aligned}$$

and note that $\mathbb{E}^* R_j^\flat(\boldsymbol{\beta}_0,\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_{j-1}) = R_j(\boldsymbol{\beta}_0,\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_{j-1})$. By the triangle inequality,

$$\begin{aligned}
\sup_{\cap_{\ell=0}^{j-1}\{\boldsymbol{\beta}_\ell\in\boldsymbol{\beta}_\ell^*+\Theta(r)\}} ||R_j^\flat(\boldsymbol{\beta}_0,\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_{j-1})||_2 \leq & \sup_{\cap_{\ell=0}^{j-1}\{\boldsymbol{\beta}_\ell\in\boldsymbol{\beta}_\ell^*+\Theta(r)\}} ||R_j(\boldsymbol{\beta}_0,\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_{j-1})||_2 \\
&+ \sup_{\cap_{\ell=0}^{j-1}\{\boldsymbol{\beta}_\ell\in\boldsymbol{\beta}_\ell^*+\Theta(r)\}} ||R_j^\flat(\boldsymbol{\beta}_0,\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_{j-1}) - R_j(\boldsymbol{\beta}_0,\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_{j-1})||_2.
\end{aligned} \tag{C.91}$$

Let $\mathscr{E}_5$ be the event that (C.6) holds for all $\ell = 0,\ldots,m-1$ with $t = 2\log n$. Then,

$\mathbb{P}(\mathscr{E}_5) \geq 1 - mn^{-2}$, and conditional on $\mathscr{E}_5$,

$$\sup_{\cap_{\ell=0}^{j-1}\{\boldsymbol{\beta}_\ell \in \boldsymbol{\beta}_\ell^* + \Theta(r)\}} ||R_j(\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{j-1})||_2 \lesssim \log\left(\tfrac{1-\tau_0}{1-\tau_j}\right) \cdot \left(m_4^{1/2} \sqrt{\frac{p+\log n}{nh}} + m_3 r + h \right) \cdot r \tag{C.92}$$

holds for all $j = 0, 1, \ldots, m$ provided that $nh \gtrsim \zeta_p^2(p+t)$. For the second term on the right-hand side of (C.91), note that

$$\begin{aligned}
&\sup_{\cap_{\ell=0}^{j-1}\{\boldsymbol{\beta}_\ell \in \boldsymbol{\beta}_\ell^* + \Theta(r)\}} ||R_j^\flat(\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{j-1}) - R_j(\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{j-1})||_2 \\
&\leq \sum_{\ell=0}^{j-1} \sup_{\boldsymbol{\beta}_\ell \in \boldsymbol{\beta}_\ell^* + \Theta(r)} \left\| \frac{1}{n} \sum_{i=1}^n e_i \big\{ \bar{K}_h(y_i - \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_\ell^*) - \bar{K}_h(y_i - \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_\ell) \big\} \boldsymbol{z}_i \right\|_2 \cdot \log\left(\tfrac{1-\tau_\ell}{1-\tau_{\ell+1}}\right) \\
&= \sum_{\ell=0}^{j-1} \Gamma_\ell(r) \cdot \log\left(\tfrac{1-\tau_\ell}{1-\tau_{\ell+1}}\right) \leq \max_{\ell=0,\ldots,j-1} \Gamma_\ell(r) \cdot \log\left(\tfrac{1-\tau_0}{1-\tau_j}\right),
\end{aligned} \tag{C.93}$$

where $\{\Gamma_\ell(r)\}_{\ell=0}^{j-1}$ are defined in (C.29).

Combining (C.91)–(C.93) proves the claimed result. □

C.5.14 Proof of Lemma C.3.6

We first show an unconditional version

$$\lim_{\delta \to 0} \limsup_{n \to \infty} \mathbb{P}\left\{ \sup_{|\tau_1 - \tau_2| < \delta} \left| \mathbb{G}_n^\flat(\tau_1) - \mathbb{G}_n^\flat(\tau_2) \right| > x \right\} = 0 \tag{C.94}$$

via the proving techniques of Lemma C.2.8. By the law of total expectation and following the arguments that lead to (C.69), it can be verified that

$$\|\mathbb{G}_n^\flat(\tau_1) - \mathbb{G}_n^\flat(\tau_2)\|_\psi \lesssim (m_4/h)^{1/2} |\tau_2 - \tau_1|,$$

for any $\tau_1 < \tau_2$ satisfying $\tau_2 - \tau_1 \geq (\zeta_p/m_4)^{1/2}(h/n)^{1/2}$, where $\psi(x) = x^2$ so that $||\cdot||_\psi$ coincides with the L_2-norm. The rest follows from the same packing argument as in the proof of Lemma C.2.8.

Then, we prove the claimed result by contradiction. Conditioned on a sequence of observed data $\{\mathbb{D}_n\}$, if (C.31) does not hold, then there is a sequence $\{\delta_m : m \subset \mathbb{N}^+\}$ such that $\lim_{m\to\infty} \delta_m = 0$, and a subsequence of natural numbers $\{n_k : k \subset \mathbb{N}^+\} \subset \mathbb{N}^+$ such that

$$\lim_{m\to\infty} \lim_{k\to\infty} \underbrace{\mathbb{P}^* \left\{ \sup_{|\tau_1-\tau_2|<\delta_m} \left| \mathbb{G}^\flat_{n_k}(\tau_1) - \mathbb{G}^\flat_{n_k}(\tau_2) \right| > x \right\}}_{:=\chi_{m,k}} > 0 \ \text{ over } \mathscr{A},$$

where $\mathscr{A}$ is an event over the data space with $\mathbb{P}(\mathscr{A}) = p_0 > 0$. This further means there are sufficiently large integers M and K such that $\chi_{m,k} \geq c_0 > 0$ over $\mathscr{A}$ for any $m \geq M$ and $k \geq K$. On the other hand, by the law of total probability,

$$\mathbb{P}\left\{ \sup_{|\tau_1-\tau_2|<\delta_m} \left| \mathbb{G}^\flat_{n_k}(\tau_1) - \mathbb{G}^\flat_{n_k}(\tau_2) \right| > x \right\} \geq \mathbb{E}\big[\chi_{m,k} \cdot \mathbb{1}\{\mathscr{A}\} \big] \geq c_0 p_0,$$

for any $m \geq M$ and $k \geq K$. So far, we have identified subsequences $\{\delta_m\}$ and $\{n_k\}$, such that $\lim_{m\to\infty} \delta_m = 0$, $\lim_{k\to\infty} n_k = \infty$, and

$$\lim_{m\to\infty} \lim_{k\to\infty} \mathbb{P}\left\{ \sup_{|\tau_1-\tau_2|<\delta_m} \left| \mathbb{G}^\flat_{n_k}(\tau_1) - \mathbb{G}^\flat_{n_k}(\tau_2) \right| > x \right\} \geq c_0 p_0 > 0,$$

which contradicts with (C.94), thus completes the proof of (C.31). □

Bibliography

ADAMCZAK, R. (2008). A tail inequality for suprema of unbounded empirical processes with applications to Markov chains. *Electron. J. Probab.* **13** 1000–1034.

ADAMCZAK, R., LITVAK, A. E., PAJOR, A. and TOMCZAK-JAEGERMANN, N. (2011). Restricted isometry property of matrices with independent columns and neighborly polytopes by random sampling. *Constr. Approx.* **34** 61–88.

AMEMIYA, T. (1982). Two stage least absolute deviations estimators. *Econometrica* **50** 689–711.

ANDERSEN, P. K., BORGAN, Ø, GILL, R. D. and KEIDING, N. (1993). *Statistical Models Based on Counting Processes*. Springer, New York.

ARCONES, M. A. (1996). The Bahadur-Kiefer representation of L_p regression estimators. *Econom. Theory* **12** 257–283.

ARCONES, M. A. and GINÉ, E. (1992). On the bootstrap of M-estimators and other statistical functionals. In *Exploring the Limits of Bootstrap* (R. LePage and L. Billard, ed.) 14–47. Wiley, New York.

BARBE, P. and BERTAIL, P. (1995). *The Weighted Bootstrap. Lecture Notes in Statistics* **98**. Springer, New York.

BARRODALE, I. and ROBERTS, F. (1974). Solution of an overdetermined system of equations in the ℓ_1 norm. *Communications of the ACM* **17** 319–320.

BARZILAI, J. and BORWEIN, J. M. (1988). Two-point step size gradient methods. *IMA J. Numer. Anal.* **8** 141–148.

BASSETT, G. and KOENKER, R. (1978). Asymptotic theory of least absolute error regression. *J. Amer. Statist. Assoc.* **73** 618–622.

BASSETT, G. and KOENKER, R. (1986). Strong consistency of regression quantiles and related empirical processes. *Econom. Theory* **2** 191–201.

BELLONI, A. and CHERNOZHUKOV, V. (2011). ℓ_1-penalized quantile regression in high-dimensional sparse models. *Ann. Statist.* **39** 82–130.

BELLONI, A., CHERNOZHUKOV, V., CHETVERIKOV, D. and FERNANDEZ-VAL, I. (2019). Conditional quantile processes based on series or many regressors. *J. Econom.* **213** 4–29.

BERTSIMAS, D., KING, A. and MAZUMDER, R. (2016). Best subset selection via a modern optimization lens. *Ann. Statist.* **44** 813–852.

BICKEL, P. J. (1975). One-step Huber estimates in the linear model. *J. Amer. Statist. Assoc.* **70** 428–434.

BOUCHERON, S., BOUSQUET, O., LUGOSI, G. and MASSART, P. (2005). Moment inequalities for functions of independent random variables. *Ann. Probab.* **33** 514–560.

BOUCHERON, S., LUGOSI, G. and MASSART, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence.* Oxford Univ. Press, Oxford.

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

BOYD, S. and VANDENBERGHE, L. (2004). *Convex Optimization.* Cambridge University Press, Cambridge.

BRADIC, J., FAN, J. and JIANG, J. (2011). Regularization for Cox's proportional hazards model with NP-dimensionality. *Ann. Statist.* **39** 3092–3120.

BUCHINSKY, M. and HAHN, J. (1998). An alternative estimator for the censored quantile regression model. *Econometrica* **66** 653–671.

CAI, T., HUANG, J. and TIAN, L. (2009). Regularized estimation for the accelerated failure time model. *Biometrics* **65** 394–404.

CHAO, S.-K., VOLGUSHEV, S. and CHENG, G. (2017). Quantile processes for semi and nonparametric regression. *Electron. J. Stat.* **11** 3272–3331.

CHATTERJEE, S. and BOSE, A. (2005). Generalized bootstrap for estimating equations. *Ann. Statist.* **33** 414–436.

CHEN, X., HONG, H. and TAROZZI, A. (2008). Semiparametric efficiency in GMM models with auxiliary data. *Ann. Statist.* **36** 808–843.

CHEN, L.-Y. and LEE, S. (2018). Exact computation of GMM estimators for instrumental

variable quantile regression models. *J. Appl. Econom.* **33** 553–567.

CHEN, X., LINTON, O. and VAN KEILEGOM, I. (2003). Estimation of semiparametric models when the criterion function is not smooth. *Econometrica* **71** 1591–1608.

CHEN, X., LIU, W. and ZHANG, Y. (2019). Quantile regression under memory constraint. *Ann. Statist.* **47** 3244–3273.

CHEN, X. and POUZO, D. (2009). Efficient estimation of semiparametric conditional moment models with possibly nonsmooth residuals. *J. Econom.* **152** 46–60.

CHEN, K., YING, Z., ZHANG, H. and ZHAO, L. (2008). Analysis of least absolute deviation. *Biometrika* **95** 107–122.

CHEN, X. and ZHOU, W.-X. (2020). Robust inference via multiplier bootstrap. *Ann. Statist.* **48** 1665–1691.

CHENG, G. and HUANG, J. Z. (2010). Bootstrap consistency for general semiparametric *M*-estimation. *Ann. Statist.* **38** 2884–2915.

CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2014). Gaussian approximation of suprema of empirical processes. *Ann. Statist.* **42** 1564–1597.

CHERNOZHUKOV, V. and HANSEN, C. (2005). An IV model of quantile treatment effects. *Econometrica* **73** 245–261.

CHERNOZHUKOV, V. and HANSEN, C. (2006). Instrumental quantile regression inference for structural and treatment effects models. *J. Econom.* **132** 491–525.

CHERNOZHUKOV, V., HANSEN, C. and WÜTHRICH, K.. (2020). Instrumental variable quantile regression. *arXiv:2009.00436*.

CHERNOZHUKOV, V. and HONG, H. (2002). Three-step censored quantile regression and extramarital affairs. *J. Am. Stat. Assoc.* **97** 872–882.

DE BACKER, M., EL GHOUCH, A. and VAN KEILEGOM, I. (2019). An adapted loss function for censored quantile regression. *J. Am. Stat. Assoc.* **114** 1126–1137.

DE BACKER, M., EL GHOUCH, A. and VAN KEILEGOM, I. (2020). Linear censored quantile regression: a novel minimum distance approach. *Scand. J. Stat.* **47** 1275–1306.

DE CASTRO, L., GALVAO, A. F., KAPLAN, D. M. and LIU, X. (2019). Smoothed GMM for quantile models. *J. Econom.* **213** 121–144.

DICKSON, E. R., GRAMBSCH, P. M., FLEMING, T. R., FISHER, L. D. and LANGWORTHY, A. (1989). Prognosis in primary biliary cirrhosis: Model for decision making. *Hepatology* **10** 1–7.

DUDEWICZ, E. J. (1992). The Generalized Bootstrap. In *Bootstrapping and Related Techniques. Lecture Notes in Economics and Mathematical Systems* **376** 31–37. Springer-Verlag, Berlin.

DUDLEY, R. M. (1979). Balls in $\boldsymbol{r}^k$ do not cut all subsets of $k+2$ points. *Adv. Math.* **31** 306–308.

EDDELBUETTEL, D. and SANDERSON, C. (2014). RcppArmadillo: Accelerating R with high-performance C++ linear algebra. *Comput. Statist. Data Anal.* **71** 1054–1063.

EFRON, B. (1967). The two sample problem with censored data. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 4: Biology and Problems of Health*, 831–853.

EFRON, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann. Statist.* **7** 1–26.

EFRON, B. and TIBSHIRANI, R. J. (1994). *An Introduction to the Bootstrap.* Chapman Hall, New York.

FALK, M. (1999). A simple approach to the generation of uniformly distributed random variables with prescribed correlations. *Comm. Statist. Simulation Comput.* **28** 785–791.

FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.* **96** 1348–1360.

FAN, J. and LI, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. *Ann. Statist.* **30** 74–99.

FAN, J., LIU, H., SUN, Q. and ZHANG, T. (2018). I-LAMM for sparse learning: Simultaneous control of algorithmic complexity and statistical error. *Ann. Statist.* **46** 814–841.

FASIOLO, M., WOOD, S. N., ZAFFRAN, M., NEDELLEC, R. and GOUDE, Y. (2020). Fast calibrated additive quantile regression. *J. Am. Stat. Assoc.* **116** 1402–1412.

FEI, Z., ZHENG, Q., HONG, H. G. and LI, Y. (2021). Inference for high dimensional censored quantile regression. *J. Am. Stat. Assoc.* https://doi.org/10.1080/01621459.2021.1957900.

FENG, X., HE, X. and HU, J. (2011). Wild bootstrap for quantile regression. *Biometrika* **98** 995–999.

FERNANDES, M., GUERRE, E. and HORTA, E. (2021). Smoothing quantile regressions. *J. Bus. Econ. Statist.* **39** 338–357.

FIRPO, S. (2007). Efficient semiparametric estimation of quantile treatment effects. *Econometrica* **75** 259–276.

FLEMING, T. R. and HARRINGTON, D. P. (1991). *Counting Processes and Survival Analysis*. Wiley, New York.

FYGENSON, M. and RITOV, Y. (1994). Monotone estimating equations for censored data. *Ann. Statist.* **22** 732–746.

GALVAO, A. F. and KATO, K. (2016). Smoothed quantile regression for panel data. *J. Econom.* **193** 92–112.

GILL, R. D. and JOHANSEN, S. (1990). A survey of product-integration with a view toward application in survival analysis. *Ann. Statist.* **18** 1501–1555.

GINÉ, E. and ZINN, J. (1990). Bootstrapping general empirical measures. *Ann. Probab.* **18** 851–869.

GU, Y., FAN, J., KONG, L., MA, S. and ZOU, H. (2018). ADMM for high-dimensional sparse regularized quantile regression. *Technometrics* **60** 319–331.

GU, S., KELLY, B. and XIU, D. (2020). Empirical asset pricing via machine learning. *Rev. Financ. Stud.* **33** 2223–2273.

GUTENBRUNNER, C. and JUREČKOVÁ, J. (1992). Regression rank scores and regression quantiles. *Ann. Statist.* **20** 305–330.

GUTENBRUNNER, C., JUREČKOVÁ, J., KOENKER, R. and PORTNOY, S. (1993). Tests of linear hypotheses based on regression rank scores. *J. Nonparametr. Stat.* **2** 307–331.

HE, X., PAN, X., TAN, K. M. and ZHOU, W.-X. (2022). Smoothed quantile regression with large-scale inference. *J. Econom.* https://doi.org/10.1016/j.jeconom.2021.07.010.

HE, X., PAN, X., TAN, K. M. and ZHOU, W.-X. (2022). Scalable estimation and inference for censored quantile regression process. *Ann. Statist.*, under revision.

HE, X. and HU, F. (2002). Markov chain marginal bootstrap. *J. Amer. Statist. Assoc.* **97** 783–795.

HE, X. and SHAO, Q.-M. (1996). A general Bahadur representation of M-estimators and its application to linear regression with nonstochastic designs. *Ann. Statist.* **24** 2608–2630.

HE, X. and SHAO, Q.-M. (2000). On parameters of increasing dimensions. *J. Mult. Anal.* **73** 120–135.

HESTENES, M. R. and STIEFEL, E. (1952). Methods of conjugate gradients for solving linear systems. *J Res. Natl. Bur. Stand.* **49** 409–436.

HONG, H. G., CHRISTIANI, D. C. and LI, Y. (2019). Quantile regression for survival data in modern cancer research: expanding statistical tools for precision medicine. *Precis. Clin. Med.* **2** 90–99.

HONORÉ, B., KHAN, S. and POWELL, J. L. (2002). Quantile regression under random censoring. *J. Econom.* **109** 67–105.

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

HOROWITZ, J. L. and LEE, S. (2007). Nonparametric instrumental variables estimation of a quantile regression model. *Econometrica* **75** 1191–1208.

HSU, D., KAKADE, S. M. and ZHANG, T. (2012). A tail inequality for quadratic forms of subgaussian random vectors. *Electron. Commun. Probab.* **17**(52): 1–6.

HU, F. and KALBFLEISCH, J. D. (2000). The estimating function bootstrap. *Canad. J. Statist.* **28** 449–499.

HUANG, Y. (2010). Quantile calculus and censored regression. *Ann. Statist.* **38** 1607–1637.

HUANG, J., MA, S. and XIE, H. (2006). Regularized estimation in the accelerated failure time model with high-dimensional covariates. *Biometrics* **62** 813–820.

HUBER, P. J. (1973). Robust estimation: Asymptotics, conjectures and Monte Carlo. *Ann. Statist.* **1** 799–821.

HUBER, P. J. (1981). *Robust Statistics*. John Wiley & Sons, New York.

JIN, Z., YING, Z. and WEI, L. J. (2001). A simple resampling method by perturbing the minimand. *Biometrika* **88** 381–390.

KAIDO, H. and WÜTHRICH, K. (2021). Decentralization estimators for instrumental variable quantile regression models. *Quantitative Economics*, to appear.

KAPLAN, D. M. and SUN, Y. (2017). Smoothed estimating equations for instrumental variables quantile regression. *Econom. Theory* **33** 105–157.

KLEY, T., VOLGUSHEV, S., DETTE, H. and HALLIN, M. (2016). Quantile spectral processes: Asymptotic analysis and inference. *Bernoulli* **22** 1770–1807.

KOCHERGINSKY, M., HE, X. and MU, Y. (2005). Practical confidence intervals for regression quantiles. *J. Comp. Graph. Statist.* **14** 41–55.

KOENKER, R. (1988). Asymptotic theory and econometric practice. *J. Appl. Econom.* **3** 139–147.

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

KOENKER, R. (2008). Censored quantile regression redux. *J. Stat. Softw.* **27**(6).

KOENKER, R. (2022). Package "quantreg", Manual: https://cran.r-project.org/web/packages/quantreg/quantreg.pdf.

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

KOENKER, R. and BASSETT, G. (1982). Tests of linear hypotheses and ℓ_1 estimation. *Econometrica* **50** 1577–1583.

KOENKER, R., CHERNOZHUKOV, V., HE, X. and PENG, L. (2017). *Handbook of Quantile Regression.* CRC Press, New York.

KOENKER, R. and D'OREY, V. (1987). Computing quantile regressions. *J. R. Statist. Soc. C* **36** 383–393.

KOENKER, R. and D'OREY, V. (1994). A remark on Algorithm AS 229: Computing dual regression quantiles and regression rank scores. *J. R. Statist. Soc. C* **43** 410–414.

KOENKER, R. and GELING, O. (2001). Reappraising medfly longevity: A quantile regression survival analysis. *J. Am. Stat. Assoc.* **96** 458–468.

KOENKER, R. and MIZERA, I. (2014). Convex optimization in R. *J. Stat. Softw.* **60**(5).

KOENKER, R. and NG, P. (2003). SparseM: A sparse matrix package for R. *J. Statist. Software* **8** 1–9.

KOENKER, R. and NG, P. (2005). A Frisch-Newton algorithm for sparse quantile regression. *Acta Mathematicae Applicatae Sinica* **21**, 225–236.

KLEINBAUM, D. G. and KLEIN, M. (2012). *Survival Analysis: A Self-Learning Text*. Springer, New York.

LANGE, K., HUNTER, D. R. and YANG, I. (2000). Optimization transfer using surrogate objective functions. *J. Comput. Graph. Stat.* **9** 1–20.

LEDOUX, M. and TALAGRAND, M. (1991). *Probability in Banach Spaces: Isoperimetry and*

Processes. Ergebnisse der Mathematik und Ihrer Grenzgebiete (3) **23**. Springer, Berlin.

LENG, C. and TONG, X. (2013). A quantile regression estimator for censored data. *Bernoulli* **19** 344–361.

LIN, D. Y., WEI, L. J. and YING, Z. (1993). Checking the Cox model with cumulative sums of martingale-based residuals. *Biometrika* **80** 557–572.

LOH, P.-L. and WAINWRIGHT, M. J. (2015). Regularized *M*-estimators with non-convexity: Statistical and algorithmic theory for local optima. *J. Mach. Learn. Res.* **16** 559–616.

MA, S. and KOSOROK, M. (2005). Robust semiparametric *M*-estimation and the weighted bootstrap. *J. Multivar. Anal.* **96** 190–217.

MAN, R., PAN, X., TAN, K. M. and ZHOU, W.-X. (2022). A unified algorithm for penalized convolution smoothed quantile regression. Preprint.

MAMMEN, E. (1989). Asymptotics with increasing dimension for robust regression with applications to the bootstrap. *Ann. Statist.* **17** 382–400.

MACHADO, J. A. F. and SANTOS SILVA, J. M. C. (2018). Quantile via moments. *J. Econom.* **213** 145–173.

MCDIARMID, C. (1989). On the method of bounded differences. In *Surveys in Combinatorics*, London Math. Soc. Lecture Note Ser., **141** 148–188. Cambridge Univ. Press, Cambridge.

NEGAHBAN, S. N., RAVIKUMAR, P., WAINWRIGHT, M. J. and YU, B. (2012). A unified framework for high-dimensional analysis of *M*-estimators with decomposable regularizers. *Statist. Sci.* **27** 538–557.

NEMIROVSKI, A. and YUDIN, D. (1983). *Problem Complexity and Method Efficiency in Optimization.* Wiley, 1983.

NEOCLEOUS, T., BRANDEN, K. V. and PORTNOY, S. (2006). Correction to censored regression quantiles by S. Portnoy, 98 (2003), 1001–1012. *J. Am. Stat. Assoc.* **101** 860–861.

ORYSHCHENKO, V. (2020). Exact mean integrated squared error and bandwidth selection for kernel distribution function estimators. *Comm. Statist. Theory Methods* **49** 1603–1628.

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

PARIKH, N. and BOYD, S. (2014). Proximal algorithms. *Found. Trends Optim.* **1** 127–239.

PARZEN, M. I., WEI, L. J. and YING, Z. (1994). A resampling method based on pivotal estimating functions. *Biometrika* **81** 341–350.

PENG, L. and HUANG, Y. (2008). Survival analysis with quantile regression models. *J. Am. Stat. Assoc.* **103** 637–649.

PENG, L. (2012). Self-consistent estimation of censored quantile regression. *J. Multivar. Anal.* **105** 368–379.

PENG, L. (2021). Quantile regression for survival data. *Annu. Rev. Stat. Appl.* **2021** 413–437.

POLLARD, D. (1991). Asymptotics for least absolute deviation regression estimators. *Econom. Theory* **7** 186–199.

PORTNOY, S. (1985). Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large; II. Normal approximation. *Ann. Statist.* **13** 1403–1417.

PORTNOY, S. (1986). On the central limit theorem in R^p when $p \to \infty$. *Probab. Theory Relat. Fields* **73** 571–583.

PORTNOY, S. (2003). Censored regression quantiles. *J. Am. Stat. Assoc.* **98** 1001–1012.

PORTNOY, S. (2012). Nearly root-n approximation for regression quantile processes. *Ann. Statist.* **40** 1714–1736.

PORTNOY, S. and KOENKER, R. (1989). Adaptive L-estimation for linear models. *Ann. Statist.* **17** 362–381.

PORTNOY, S. and KOENKER, R. (1997). The Gaussian hare and the Laplacian tortoise: Computability of squared-error versus absolute-error estimators. *Statist. Sci.* **12** 279–300.

PORTNOY, S. and LIN, G. (2010). Asymptotics for censored regression quantiles. *J. Nonparametr. Stat.* **22** 115–130.

POWELL, J. L. (1984). Least absolute deviations estimation for the censored regression model. *J. Econom.* **25** 303–325.

POWELL, J. L. (1986). Censored regression quantiles. *J. Econom.* **32** 143–155.

PRAESTGAARD, J. and WELLNER, J. (1993). Exchangeably weighted bootstraps of the general empirical process. *Ann. Probab.* **21** 2053–2086.

RUPPERT, D. and CARROLL, R. J. (1980). Trimmed least squares estimation in the linear model. *J. Amer. Statist. Assoc.* **75** 828–838.

SHEDDEN, K., TAYLOR, J. M., ENKEMANN, S. A., ..., JACOBSON, J. W. and BEER, D. G. (2008). Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. *Nat. Med.* **14** 822–827.

SHERWOOD, B. and MAIDMAN, A. (2022). Package 'rqPen'. Reference manual: https://cran.r-project.org/web/packages/rqPen/rqPen.pdf.

SHOWS, J. H., LU, W and ZHANG, H. H. (2010). Sparse estimation and inference for censored median regression. *J. Statist. Plann. Inference* **140** 1903–1917.

SILVERMAN, B. (1986). *Density Estimation for Statistics and Data Analysis.* Chapman Hall, New York.

SPOKOINY, V. (2013). Bernstein–von Mises theorem for growing parameter dimension. Available at *arXiv:1302.3430*.

SPOKOINY, V. and ZHILOVA, M. (2015). Bootstrap confidence sets under model misspecification. *Ann. Statist.* **43** 2653–2675.

SUN, J. H., PENG, L, HUANG, Y. and LAI, H. J. (2016). Generalizing quantile regression for counting processes with applications to recurrent events. *J. Am. Stat. Assoc.* **111** 145–156.

SUN, Q., ZHOU, W.-X. and FAN, J. (2020). Adaptive Huber regression. *J. Amer. Statist. Assoc.* **115** 254–265.

TAN, K. M., WANG, L. and ZHOU, W.-X. (2022). High-dimensional quantile regression: Convolution smoothing and concave regularization. *J. Roy. Statist. Soc. Ser. B* **84**(1) 205–233.

THERNEAU, T. M., GRAMBSCH, P. M. and FLEMING, T. R. (1990). Martingale-based residuals for survival models. *Biometrika* **77** 147–160.

TIAN, L., ZUCKER, D. and WEI, L. J. (2005). On the Cox model with time-varying regression coefficients. *J. Am. Stat. Assoc.* **100** 172–183.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc.* B **58** 267–288.

TYURIN, I. S. (2011). On the convergence rate in Lyapunov's theorem. *Theory Probab. Appl.* **55** 253–270.

VAN DE GEER, S. (2000). *Empirical Processes in M-Estimation*. Cambridge Univ. Press, Cambridge.

VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical*

Processes: With Applications to Statistics. Springer, New York.

VERSHYNIN, R. (2012). Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing* (Y. Eldar and G. Kutyniok, eds.) 210–268. Cambridge Univ. Press, Cambridge.

VERSHYNIN, R. (2018). *High-Dimensional Probability: An Introduction with Applications in Data Science.* Cambridge University Press, Cambridge.

VOLGUSHEV, S, VAGENER, J. and DETTE, H. (2014). Censored quantile regression processes under dependence and penalization. *Electron. J. Stat.* **8** 2405–2447.

WAINWRIGHT, M. J. (2019) *High-Dimensional Statistics: A Non-Asymptotic Viewpoint.* Cambridge Univ. Press, Cambridge.

WAND, M. P. and SCHUCANY, W. R. (1990). Gaussian-based kernels. *Canad. J. Statist.* **18** 197–204.

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

WANG, H. J. and WANG, L. (2009). Locally weighted censored quantile regression. *J. Am. Stat. Assoc.* **104** 1117–1128.

WANG, L., WU, Y. and LI, R. (2012). Quantile regression for analyzing heterogeneity in ultra-high dimension. *J. Amer. Statist. Assoc.* **107** 214–222.

WANG, H. J., ZHOU, J. and LI, Y. (2013). Variable selection for censored quantile regression. *Statist. Sinica* **23** 145–167.

WELLNER, J. A. and ZHAN, Y. (1996). Bootstrapping Z-estimators. *Technical report*. Department of Statistics, University of Washington, Seattle.

WELSH, A. H. (1989). On M-processes and M-estimation. *Ann. Statist.* **15** 337–361.

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

WU, Y., MA, Y. and YIN, G. (2015). Smoothed and corrected score approach to censored quantile regression with measurement errors. *J. Am. Stat. Assoc.* **110** 1670–1683.

YANG, X., NARISETTY, N. N. and HE, X. (2018). A new approach to censored quantile regression estimation. *J. Comput. Graph. Stat.* **27** 417–425.

YING, Z., JUNG, S. H. and WEI, L. J. (1995). Survival analysis with median regression models. *J. Am. Stat. Assoc.* **90** 178–184.

YU, L., LIN, N. and WANG, L. (2017). A parallel algorithm for large-scale nonconvex penalized quantile regression. *J. Comput. Graph. Stat.* **26** 935–939.

ZHANG, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38** 894–942.

ZHAO, L. C., RAO, C. R. and CHEN, X. R. (1993). A note on the consistency of *M*-estimates in linear models. In *Stochastic Processes: A Festschrift in Honour of Gopinath Kallianpur*, 359–367. Springer, New York.

ZHENG, Q., PENG, L. and HE, X. (2015). Globally adaptive quantile regression with ultra-high dimensional data. *Ann. Statist.* **43** 2225–2258.

ZHENG, Q., PENG, L. and HE, X. (2018). High dimensional censored quantile regression. *Ann. Statist.* **46** 308–343.

ZHU, Y. (2018). k-step correction for mixed integer linear programming: a new approach for instrumental variable quantile regressions and related problems. *arXiv:1805.06855.*

ZOU, H. (2006). The adaptive lasso and its oracle properties. *J. Am. Stat. Assoc.* **101** 1418–1429.

ZOU, H. and LI, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Statist.* **36** 1509–1533.