Title

Multiplier bootstrap for quantile regression: non-asymptotic theory under random design

Journal

Information and Inference A Journal of the IMA, 10(3)

ISSN

2049-8764

Authors

Pan, Xiaoou
Zhou, Wen-Xin

Publication Date

2021-09-14

DOI

10.1093/imaiai/iaaa006

Peer reviewed

Information and Inference: A Journal of the IMA (2021) **10**, 813–861
https://doi.org/10.1093/imaiai/iaaa006
Advance Access publication on 11 May 2020

Multiplier bootstrap for quantile regression: non-asymptotic theory under random design ®

Xiaoou Pan
Department of Mathematics, University of California, San Diego, La Jolla, CA 92093, USA
Email: xip024@ucsd.edu

and

Wen-Xin Zhou†
Department of Mathematics, University of California, San Diego, La Jolla, CA 92093, USA
†Corresponding author. Email: wez243@ucsd.edu

[Received on 7 August 2019; revised on 18 February 2020; accepted on 13 March 2020]

This paper establishes non-asymptotic concentration bound and Bahadur representation for the quantile regression estimator and its multiplier bootstrap counterpart in the random design setting. The non-asymptotic analysis keeps track of the impact of the parameter dimension d and sample size n in the rate of convergence, as well as in normal and bootstrap approximation errors. These results represent a useful complement to the asymptotic results under fixed design and provide theoretical guarantees for the validity of Rademacher multiplier bootstrap in the problems of confidence construction and goodness-of-fit testing. Numerical studies lend strong support to our theory and highlight the effectiveness of Rademacher bootstrap in terms of accuracy, reliability and computational efficiency.

Keywords: quantile regression; multiplier bootstrap; robustness; concentration inequality; Bahadur representation; confidence interval; goodness-of-fit test.

1. Introduction

1.1 *Quantile regression*

Since Koenker and Bassett's (1978) seminal work, quantile regression has attracted enormous attention in statistics, econometrics and related fields primarily due to two advantages over the (conditional) mean regression: (i) robustness against outliers in the response or heavy-tailed errors and (ii) the ability to explore heterogeneity in the response that are associated with the covariates. We refer to the monograph by Koenker (2005) for an overview of the statistical theory and methods and computational aspects of quantile regression.

Classical theory of quantile regression includes statistical consistency (see, e.g. Zhao *et al.*, 1993, for weak consistency and Bassett & Koenker, 1986, for strong consistency), asymptotic normality (Bassett & Koenker, 1978; Pollard, 1991) and Bahadur representation (Portnoy & Koenker, 1989; He & Shao, 1996; Arcones, 1996). 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 and satisfy some (asymptotic and non-asymptotic) conditions and the only randomness arises from the regression

® Symbol indicates reproducible data.

errors $\{\varepsilon_i\}_{i=1}^n$. A comprehensive review of the asymptotic theory under fixed design can be found in Sections 4.1–4.3 of Koenker (2005).

In contrast to fixed designs, more recent work in statistics has emphasized non-asymptotic results in the random design setting, where the covariates $\{\boldsymbol{x}_i\}_{i=1}^n$ are treated as random vectors (Hsu *et al.*, 2014; Wainwright, 2019). 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. As stated in Hsu *et al.* (2014), a major difference between fixed and random designs is that the fixed design setting does not directly address out-of-sample prediction. Specifically, a fixed design analysis assesses the accuracy of the estimator on the observed data, while the predictive performance on unseen data is of primary concern of a random design analysis. Even though extensive studies have been carried out on ordinary and regularized least squares estimators (Hsu *et al.*, 2014; Wainwright, 2019), it is not naturally clear whether similar results remain valid for quantile regression. 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 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.

In Section 2.1 of this paper, we will establish non-asymptotic concentration bound (Theorem 2.1) and Bahadur representation (Theorem 2.2) of the quantile regression estimator under mild conditions on the random predictor and noise variable. To prove Theorem 2.1, we propose a new device to prove a local *restricted strong convexity* (RSC) property of the empirical quantile loss, see Proposition 4.2. The notion of RSC was introduced by Negahban *et al.* (2012) to analyze convex regularized M-estimators and extended by Loh & Wainwright (2015) to the case of nonconvex functions. Thus far the RSC property has only been established for locally strongly convex and twice differentiable loss functions (Loh & Wainwright, 2015; Pan *et al.*, 2019). New techniques are therefore required to deal with piecewise linear functions, typified by the quantile loss and hinge loss. The proof of Theorem 2.2, the Bahadur representation, builds on the concentration bound in Theorem 2.1 along with techniques from empirical process theory. These results are non-asymptotic with explicit errors, which allow to track the impact of the parameter dimension d and of the sample size n in quantile regression. These non-asymptotic results, to the best of our knowledge, are new to the previous asymptotic results under fixed designs.

1.2 *Statistical inference for quantile regression*

In addition to the finite sample theory of standard quantile regression, we are also interested in two fundamental statistical inference problems: (i) the construction of confidence intervals and (ii) goodness-of-fit test. 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 $1/f_{\varepsilon|\boldsymbol{x}}(0)$, also known as the sparsity, where $f_{\varepsilon|\boldsymbol{x}}(\cdot)$ is the conditional density function of ε given $\boldsymbol{x}$, or the regression rank scores (Gutenbrunner & Jurečková, 1992). Resampling, or bootstrap calibration, methods (Efron, 1979) are commonly used for quantile regression inference because they are more robust against heteroscedastic errors and bypass the estimation of sparsity although at the cost of computing time. Over the past two decades, various bootstrap calibration methods have been developed for constructing confidence intervals, including the residual bootstrap and pairwise bootstrap (see Section 9.5 of Efron & Tibshirani, 1994), bootstrapping pivotal estimation functions method (Parzen *et al.*, 1994), Markov chain marginal bootstrap (He & Hu, 2002; Kocherginsky *et al.*, 2005) and wild bootstrap (Feng *et al.*, 2011). 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 paper 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 & Zinn (1990), Arcones & Giné (1992), Praestgaard & Wellner (1993) and Wellner & Zhan (1996), among others. Rigorous theoretical guarantees of the multiplier bootstrap for M-estimation can be found in Chatterjee & Bose (2005) and Ma & Kosorok (2005), in which $\sqrt{n}$-consistency and asymptotic normality are established. See also Cheng & 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 & Zhilova (2015) for the use of multiplier bootstrap on constructing likelihood-based confidence sets and Chen & Zhou (2019) for a systematic study of multiplier bootstrap for adaptive Huber regression (Sun *et al.*, 2019) with applications to large-scale multiple testing for heavy-tailed data.

As stated in the previous section, the major theoretical challenge arises from the lack of smoothness and strong convexity of the quantile loss. New techniques are in demand. In Section 2.2, we will first revisit the multiplier bootstrap in the problem of confidence estimation for quantile regression. Next, we will provide new non-asymptotic theory for bootstrap estimators, including the conditional deviation bound (Theorem 2.4) and Bahadur representation (Theorem 2.5) conditioned on data already seen. We justify the validity of the multiplier bootstrap via a distributional approximation result (Theorem 2.6), which characterizes the difference in distribution between the regression estimator and its bootstrap counterpart. In Section 2.3, we further discuss the use of multiplier bootstrap on goodness-of-fit testing, extending the special case of median regression studied by Chen *et al.* (2008).

1.3 *Notation*

Let us summarize our notation. For every integer $k \geqslant 1$, we use $\mathbb{R}^k$ to denote the k-dimensional Euclidean space. The inner product of any two vectors $\boldsymbol{u} = (u_1, \ldots, u_k)^\intercal, \boldsymbol{v} = (v_1, \ldots, v_k)^\intercal \in \mathbb{R}^k$ is defined by $\boldsymbol{u}^\intercal \boldsymbol{v} = \langle \boldsymbol{u}, \boldsymbol{v} \rangle = \sum_{i=1}^k u_i v_i$. We use $\| \cdot \|_p$ $(1 \leqslant p \leqslant \infty)$ to denote the ℓ_p-norm in $\mathbb{R}^k$: $\|\boldsymbol{u}\|_p = (\sum_{i=1}^k |u_i|^p)^{1/p}$ and $\|\boldsymbol{u}\|_\infty = \max_{1 \leqslant i \leqslant k} |u_i|$. For $k \geqslant 2$, $\mathbb{S}^{k-1} = \{\boldsymbol{u} \in \mathbb{R}^k : \|\boldsymbol{u}\|_2 = 1\}$ denotes the unit sphere in $\mathbb{R}^k$.

Throughout this paper, we use bold capital letters to represent matrices. For $k \geqslant 2$, $\mathbf{i}_k$ 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 $r \geqslant 0$, define the Euclidean ball and ellipse as $\mathbb{B}^k(r) = \{\boldsymbol{u} \in \mathbb{R}^k : \|\boldsymbol{u}\|_2 \leqslant r\}$ and $\mathbb{B}_{\mathbf{A}}(r) = \{\boldsymbol{u} \in \mathbb{R}^k : \|\boldsymbol{u}\|_{\mathbf{A}} \leqslant r\}$, respectively. For any integer $d \geqslant 1$, we write $[d] = \{1, \ldots, d\}$. For any set $\mathcal{S}$, we use $|\mathcal{S}|$ to denote its cardinality, i.e. the number of elements in $\mathcal{S}$.

2. Random design quantile regression

2.1 *Finite sample theory under random design*

We consider a response variable y and d-dimensional covariates $\boldsymbol{x} = (x_1, \ldots, x_d)^\intercal$ 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 $\boldsymbol{\beta}^* = (\beta_1^*, \ldots, \beta_d^*)^\intercal \in \mathbb{R}^d$. Here we assume $x_1 \equiv 1$ so that β_1^* represents the intercept. Let $\{(y_i, \boldsymbol{x}_i)\}_{i=1}^n$ be

TABLE 1 *Summary of scaling conditions required for normal approximation under the Huber and pinball loss functions*

Loss function	Design	Scaling condition
Huber loss (Portnoy, 1985)	Mixed Gaussian (with symmetric noise)	$(d \log n)^{3/2} = o(n)$
Huber loss (Portnoy, 1986)	Fixed design (with symmetric noise)	$d^2 = o(n)$
Huber loss (Chen & Zhou, 2019)	Sub-Gaussian (with asymmetric noise)	$d^2 = o(n)$
Pinball loss (Welsh, 1989)	Fixed design	$d^3(\log n)^2 = o(n)$
Pinball loss (this work)	Sub-Gaussian	$d^3(\log n)^2 = o(n)$

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 $\mathbb{P}(\varepsilon_i \leqslant 0 \,|\, \boldsymbol{x}_i) = \tau$. The quantile regression estimator of $\boldsymbol{\beta}^*$ is then defined as

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

where

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

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. We refer to Chapter 4 of Koenker (2005) for the classical fixed design and asymptotic analysis of quantile regression. See also Remark 2.2 and Table 1 below for a comparison of quantile regression and smooth robust regression in terms of the scalings of the pair (n, d).

First, we specify the conditions on the random pair $(\boldsymbol{x}, \varepsilon)$ under which the analysis applies.

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

Condition 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 $\mathbb{R}^d$ with center at the origin and radius $\sqrt{d}$, (iii) uniform distribution on the Euclidean ball and (iv) uniform distribution on the unit cube $[-1, 1]^d$. 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 (Regularity condition on error distribution). Let $f_{\varepsilon|\boldsymbol{x}}(\cdot)$ be the conditional probability density function of ε given $\boldsymbol{x}$, which is continuous on its support. Moreover, there exist constants

$\bar{f} \geqslant \underline{f} > 0$ and $L_0 > 0$ such that

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

Condition 2 on the conditional density function of ε given $\boldsymbol{x}$ is standard and routinely used in the study of quantile regression.

Throughout this paper, '$\lesssim$' stands for '$\leqslant$' up to constants that are independent of (n, d) but may depend on the constants in Conditions 1 and 2. Our first main result characterizes the non-asymptotic deviation of the quantile regression estimator.

THEOREM 2.1 Assume Conditions 1 and 2 hold. Then, for any $t \geqslant 0$, the quantile regression estimator $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\tau)$ $(0 < \tau < 1)$ given in (2.2) satisfies

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma}} \leq \frac{c_1}{\underline{f}}\sqrt{\frac{d+t}{n}} \tag{2.4}$$

with probability at least $1-2e^{-t}$ as long as $n \geqslant c_2 L_0^2 \underline{f}^{-4}(d+t)$, where $c_1, c_2 > 0$ are constants depending only on υ_0.

The following theorem provides a non-asymptotic version of the Bahadur representation for the quantile regression estimator see Section 4.3 in Koenker (2005).

THEOREM 2.2 Suppose that, in addition to the conditions in Theorem 2.1, $\sup_{u\in\mathbb{R}} |f_{\varepsilon|\boldsymbol{x}}(u)| \leqslant M_0$ almost surely for some $M_0 > 0$. Then, for any $t \geqslant 0$,

$$\begin{aligned}&\left\|\mathbf{S}^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) + \mathbf{S}^{-1/2}\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\{I(\varepsilon_i \leqslant 0) - \tau\}\right\|_2 \\ &\leqslant c_3\left\{\frac{(d+t)^{1/4}(d\log n + t)^{1/2}}{n^{3/4}} + (d+\log n)^{1/2}\frac{d\log n}{n} + (d\log n)^{1/2}\frac{t}{n}\right\}\end{aligned} \tag{2.5}$$

with probability at least $1 - 4e^{-t}$ whenever $n \geqslant c_2 L_0^2 \underline{f}^{-4}(d+t)$, where $\mathbf{S} = \mathbb{E}\{f_{\varepsilon|\boldsymbol{x}}(0)\boldsymbol{x}\boldsymbol{x}^\intercal\}$, and $c_3 > 0$ is a constant depending only on $(\upsilon_0, \underline{f}, \bar{f}, L_0, M_0)$.

REMARK 2.1 With some basic analysis, the property that $\sup_{u\in\mathbb{R}} |f_{\varepsilon|\boldsymbol{x}}(u)| \leqslant M_0$ almost surely is a consequence of Condition 2 with M_0 depending implicitly on $(\bar{f}, L_0)$. Hence, introducing the constant M_0 is not to initiate an additional assumption but to simplify the theorem and its proof.

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 2.3 Let $\boldsymbol{\lambda} \in \mathbb{R}^d$ be a deterministic vector that defines a linear contrast of interest. Under the conditions of Theorem 2.2, it holds that

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

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 2.2 (Large-d asymptotics). A broader view of classical asymptotics recognizes that the parametric dimension of appropriate model sequences may tend to infinity with the sample size, that is $d = d_n \to \infty$ as $n \to \infty$. Such considerations, however, are rarely found in the quantile regression literature. In the standard quantile regression setting, Welsh (1989) shows that $d^3(\log n)^2/n \to 0$ suffices for a normal approximation, which provides some support to the viability of observed rates of parametric growth in the applied literature (Koenker, 1988).

In the (sub-Gaussian) random design setting, the obtained non-asymptotic Bahadur representation (2.5) with $t = \log n$ reads:

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

Combined with a multivariate central limit theorem (Portnoy, 1986) or Theorem 2.3, this shows that the normal approximation holds as long as $d^3(\log n)^2/n \to 0$, which matches the scaling under fixed design although the proofs are entirely different. For smooth robust regression estimators, the scaling conditions required for asymptotic normality can be weakened. A prototypical example is Huber's M-estimator. Note that the Huber loss has an absolutely continuous derivative and is twice differentiable except at two points. Portnoy (1985) obtains the scaling condition $(d\log n)^{3/2}/n \to 0$ that validates asymptotic normality when the predictors $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ form a sample from a mixed multivariate normal distribution in $\mathbb{R}^d$. In the case of random, non-Gaussian predictors and of symmetric noise, d^2/n is necessary for normal approximation, see Portnoy (1985, 1986).

2.2 *Multiplier bootstrap and confidence estimation*

Let $\mathcal{R}_n = \{e_1, \ldots, e_n\}$ be a sequence of independent Rademacher random variables that are independent of the observed data $\mathcal{D}_n = \{(y_i, \boldsymbol{x}_i)\}_{i=1}^n$. Specifically, $e_i \in \{-1, 1\}$ 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

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

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

Let $\mathbb{E}^*(\cdot) = \mathbb{E}(\cdot\,|\mathcal{D}_n)$ and $\mathbb{P}^*(\cdot) = \mathbb{P}(\cdot\,|\mathcal{D}_n)$ be the conditional expectation and probability given $\mathcal{D}_n$, respectively. Then we have $\mathbb{E}^*\{Q_n^\flat(\boldsymbol{\beta})\} = Q_n(\boldsymbol{\beta})$ for any $\boldsymbol{\beta} \in \mathbb{R}^d$. This indicates that the quantile estimator $\hat{\boldsymbol{\beta}}(\tau) = (\hat{\beta}_1, \ldots, \hat{\beta}_d)^\intercal$ in the $\mathcal{D}_n$-world is the target parameter in the bootstrap world:

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

This simple observation motivates the following multiplier bootstrap estimator:

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

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

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

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

Efron's percentile interval is of the form

$$\mathcal{I}_j^{\mathrm{per}} = [\hat{\zeta}_{j,1-\alpha/2},\ \hat{\zeta}_{j,\alpha/2}],\ \ j = 1, \ldots, d. \tag{2.10}$$

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

$$\mathcal{I}_j^{\mathrm{norm}} = [\hat{\beta}_j - z_{\alpha/2}\hat{\mathrm{se}}_j^{\mathrm{boot}},\ \hat{\beta}_j + z_{\alpha/2}\hat{\mathrm{se}}_j^{\mathrm{boot}}],\ \ j = 1, \ldots, d, \tag{2.11}$$

where $\hat{\mathrm{se}}_j^{\mathrm{boot}}$ is the conditional standard deviation of $\hat{\beta}_j^\flat$ given $\mathcal{D}_n$, 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 $\hat{\boldsymbol{\beta}}^\flat(\tau) - \hat{\boldsymbol{\beta}}(\tau)$ to approximate the distribution of the pivot $\hat{\boldsymbol{\beta}}(\tau) - \boldsymbol{\beta}^*$, is the pivotal interval. Specifically, the $1-\alpha$ bootstrap pivotal confidence intervals for β_j^*'s are

$$\mathcal{I}_j^{\mathrm{piv}} = [2\hat{\beta}_j - \hat{\zeta}_{j,\alpha/2},\ 2\hat{\beta}_j - \hat{\zeta}_{j,1-\alpha/2}],\ \ j = 1, \ldots, d. \tag{2.12}$$

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 $\hat{\beta}_j$.

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

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

where $\boldsymbol{r}_n$ is the higher-order remainder term. Heuristically, the bootstrap estimator $\hat{\boldsymbol{\beta}}^\flat(\tau)$ can be viewed as the quantile regression estimator of $\hat{\boldsymbol{\beta}}(\tau)$ in the bootstrap world under the model $y_i = \langle \boldsymbol{x}_i, \hat{\boldsymbol{\beta}}(\tau)\rangle + \varepsilon_i^\flat$. According to the Bahadur representation, it can be written as $y_i \approx \langle \boldsymbol{x}_i, \boldsymbol{\beta}^*\rangle + (1/n)\sum_{i=1}^n \langle \boldsymbol{x}_i, \mathbf{S}^{-1}\boldsymbol{x}_i\rangle\{\tau -$

$I(\varepsilon_i \leqslant 0)\}$. The accuracy of the percentile interval, however, relies on the property that $\hat{\boldsymbol{\beta}}^\flat_\tau$ is randomly concentrated around $\boldsymbol{\beta}^*$. Motivated by this observation and the finite-sample correction method used in Feng *et al.* (2011), for practical implementation we replace the original response y_i in the multiplier bootstrap by $\hat{y}_i = y_i - \{\hat{f}_\varepsilon(0)\}^{-1} h_i\{\tau - I(\hat{\varepsilon}_i \leqslant 0)\}$, where $h_i = \boldsymbol{x}_i^\mathsf{T}(\sum_{j=1}^n \boldsymbol{x}_j\boldsymbol{x}_j^\mathsf{T})^{-1}\boldsymbol{x}_i$ and $\hat{f}_\varepsilon(0)$ is estimated from the fitted residuals $\hat{\varepsilon}_i = y_i - \langle \boldsymbol{x}_i, \hat{\boldsymbol{\beta}}(\tau)\rangle$. In particular, the density estimate $\hat{f}_\varepsilon$ employs the adaptive kernel method (Silverman, 1986), which is implemented in the quantreg package as function akj (Koenker, 2019).

Back to $\hat{\boldsymbol{\beta}}^\flat$ defined in (2.8), the following result provides a conditional deviation inequality, conditioned on some event that occurs with high probability.

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

$$\|\hat{\boldsymbol{\beta}}^\flat - \boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma}} \leqslant c_4\sqrt{\frac{d+t}{n}} \tag{2.13}$$

as long as $n \geqslant c_5(d+t)$, where $c_4, c_5 > 0$ are constants depending only on $(\upsilon_0, \underline{f}, L_0)$.

To characterize the distribution of $\hat{\boldsymbol{\beta}}^\flat$ conditional on the initial sample $\mathcal{D}_n = \{(y_i, \boldsymbol{x}_i)\}_{i=1}^n$, we establish in the following result a conditional Bahadur representation under $\mathbb{P}^*$.

THEOREM 2.5 Suppose that the conditions in Theorem 2.2 hold. Under the scaling $n \gtrsim d + \log n$, there exists some event $\mathcal{E}_n$ with $\mathbb{P}(\mathcal{E}_n) \geqslant 1 - 4n^{-1}$ such that, with $\mathbb{P}^*$-probability at least $1 - n^{-1}$ conditioned on $\mathcal{E}_n$,

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

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 $\chi_n = O_{\mathbb{P}}\{(d+\log n)^{1/4}(d\log n)^{1/2}n^{-3/4} + (d+\log n)^{1/2}d\log(n)\,n^{-1}\}$.

We end this section with a distributional approximation result, which establishes the validity of the (Rademacher) multiplier bootstrap for approximating the distributions of linear contrasts of the quantile regression estimator.

THEOREM 2.6 Let $\boldsymbol{\lambda} \in \mathbb{R}^d$ be an arbitrary d-vector defining a linear contrast of interest. Assume Conditions 1 and 2 hold, and that the parameter dimension $d = d_n$, as a function of the sample size n, satisfies the scaling $d^3(\log n)^2 = o(n)$. Then, as $n \to \infty$,

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

2.3 *Goodness-of-fit testing*

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

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

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

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

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 $\hat{\boldsymbol{\beta}}^\flat \in \operatorname{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^d} Q_n^\flat(\boldsymbol{\beta})$ and $\hat{\boldsymbol{\beta}}_0^\flat \in \operatorname{argmin}_{\boldsymbol{\beta}\in\Omega_0} Q_n^\flat(\boldsymbol{\beta})$ be the bootstrap statistics in the full model and null model, respectively. Motivated by Chen *et al.* (2008), we consider the bootstrap test statistic

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

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,

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

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

It is worth noticing that the above method was first proposed and studied by Chen *et al.* (2008) using standard exponential weights in the case of median regression and can be implemented by the R package quantreg (Koenker, 2019). As discussed earlier, the Rademacher multiplier bootstrap is computationally more attractive and also has provable finite-sample guarantees. See Sections 3.2 and B.2 for a thorough numerical comparison.

3. 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.

3.1 *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 *et al.* (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 & Tibshirani, 1994);
- pwy: a resampling method based on pivotal estimating functions (Parzen *et al.*, 1994);
- wild: wild bootstrap with Rademacher weights (Feng *et al.*, 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 (Koenker, 2019).

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):

$$y_i = \beta_0^* + \langle \boldsymbol{x}_i, \boldsymbol{\beta}^* \rangle + \varepsilon_i, \quad i = 1, \ldots, n. \tag{3.1}$$

2. (Heteroscedastic model):

$$y_i = \beta_0^* + \langle \boldsymbol{x}_i, \boldsymbol{\beta}^* \rangle + \frac{2\exp(x_{i1})}{1 + \exp(x_{i1})}\varepsilon_i, \quad i = 1, \ldots, n. \tag{3.2}$$

Here we use separate notations to differentiate the intercept β_0^* and coefficient vector $\boldsymbol{\beta}^* \in \mathbb{R}^d$. 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: $\boldsymbol{x}_i \sim \mathcal{N}(0, \mathbf{i}_d)$ for $i = 1, \ldots, n$.
2. Weakly correlated design: first generate a covariance matrix $\boldsymbol{\Sigma} = (\sigma_{jk})_{1 \leqslant j,k \leqslant d}$ with diagonal entries σ_{jj} independently drawn from 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 $\boldsymbol{\Sigma} = (\sigma_{jk})_{1 \leqslant j,k \leqslant d}$ 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 $\beta_0^* = 2$, $\boldsymbol{\beta}^* = (2, \ldots, 2)^\intercal$ and $(n, d) = (200, 10)$. 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,

TABLE 2 *Average coverage probabilities and confidence interval (CI) widths over all the coefficients under homoscedastic model (3.1) 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

TABLE 3 *Average coverage probabilities and CI widths over all the coefficients under heteroscedastic model (3.2) 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 4 *Average coverage probabilities and CI widths (in brackets) over all the coefficients under homoscedastic model (3.1) 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)

3 and B7– B10 in Section B.1 of the Appendix display the average coverage probabilities and average interval widths over all the regression coefficients based on 200 Monte Carlo simulations.

From Tables 2, 3 and B7– B10 (in the Appendix), 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. 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 4. After taking into account the interval width, we recommend the multiplier bootstrap percentile method that has the best overall performance.

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.

3.2 *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 & 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 (Koenker, 2019).

We generate data vectors the same way as in Section 3.1. Moreover, we set $(n, d) = (200, 15)$, 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 5, 6 and B11–B14 in Section B.2 of the Appendix display the average type I error and power over 200 Monte Carlo simulations.

From Tables 5 and 6, 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 1 is a visualization of Tables 5 and 6 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.

4. 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.

4.1 *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 $Q_n : \mathbb{R}^d \to \mathbb{R}$ is convex, we define its subdifferential ∂Q_n by

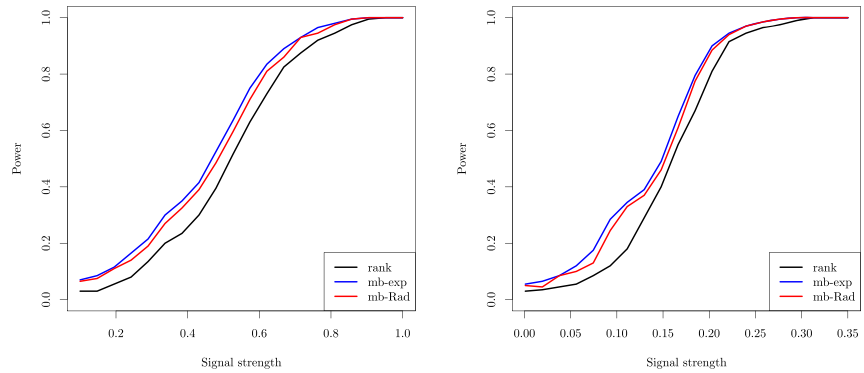
$$\partial Q_n(\boldsymbol{\beta}) = \big\{ \boldsymbol{\xi} \in \mathbb{R}^d : Q_n(\boldsymbol{\beta}') \geqslant Q_n(\boldsymbol{\beta}) + \langle \boldsymbol{\xi}, \boldsymbol{\beta}' - \boldsymbol{\beta} \rangle \text{ for all } \boldsymbol{\beta} \in \mathbb{R}^d \big\}. \tag{4.1}$$

TABLE 5 *Average type I error and power under homoscedastic model (3.1) with type I mixture normal error*

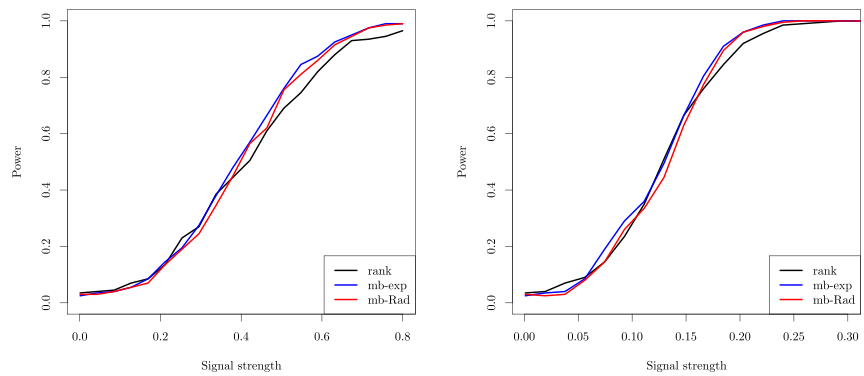
Independent Gaussian design												
	Type I error under null model				Power under sparse model				Power under dense model			
α	Wald	rank	mb-exp	mb-Rad	Wald	rank	mb-exp	mb-Rad	Wald	rank	mb-exp	mb-Rad
0.01	0.370	0.000	0.000	0.005	0.805	0.185	0.295	0.330	0.580	0.035	0.045	0.075
0.05	0.490	0.025	0.055	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 6 *Average type I error and power under heteroscedastic model (3.2) 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



(a) Homoscedastic model (3.1) with sparse signal. (b) Homoscedastic model (3.1) with dense signal.



(c) Heteroscedastic model (3.2) with sparse signal. (d) Heteroscedastic model (3.2) with dense signal.

FIG. 1. Power curves of the three methods under independent design and type I mixture normal error with $\alpha = 0.05$.

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 $\boldsymbol{\xi}_{\boldsymbol{\beta}} = (\xi_{\boldsymbol{\beta},1}, \ldots, \xi_{\boldsymbol{\beta},d})^{\mathsf{T}}$ satisfying, for $j = 1 \ldots, d$,

$$\begin{aligned}\xi_{\boldsymbol{\beta},j} = &-\frac{\tau}{n}\sum_{i=1}^{n} x_{ij} I(y_i > \langle \boldsymbol{x}_i, \boldsymbol{\beta}\rangle) \\ &+ \frac{1-\tau}{n}\sum_{i=1}^{n} x_{ij} I(y_i < \langle \boldsymbol{x}_i, \boldsymbol{\beta}\rangle) - \frac{1}{n}\sum_{i=1}^{n} x_{ij} v_i I(y_i = \langle \boldsymbol{x}_i, \boldsymbol{\beta}\rangle),\end{aligned} \tag{4.2}$$

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

Of particular interest is the subdifferential $\partial Q_n(\boldsymbol{\beta}^*)$ under model (2.1). By (4.2), every vector $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_d)^\intercal \in \partial Q_n(\boldsymbol{\beta}^*)$ can be written as

$$\xi_j = -\frac{\tau}{n}\sum_{i=1}^{n} x_{ij}\{I(\varepsilon_i > 0) - (1-\tau)\} + \frac{1-\tau}{n}\sum_{i=1}^{n} x_{ij}\{I(\varepsilon_i < 0) - \tau\} - \frac{1}{n}\sum_{i=1}^{n} x_{ij} v_i I(\varepsilon_i = 0), \ \ j = 1, \ldots, d, \tag{4.3}$$

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

PROPOSITION 4.1 Assume Conditions 1 and 2 hold. Then, every subgradient $\boldsymbol{\xi}_{\boldsymbol{\beta}^*} \in \partial Q_n(\boldsymbol{\beta}^*)$ satisfies

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

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

PROPOSITION 4.2 Assume Conditions 1 and 2 hold. Then, for any $t \geqslant 0$, it holds with probability at least $1 - e^{-t}/2$ that

$$\langle \boldsymbol{\xi}_{\boldsymbol{\beta}} - \boldsymbol{\xi}_{\boldsymbol{\beta}^*}, \boldsymbol{\beta} - \boldsymbol{\beta}^* \rangle \geqslant \frac{1}{8} \underline{f} \, \| \boldsymbol{\beta} - \boldsymbol{\beta}^* \|_{\boldsymbol{\Sigma}}^2 - 4\upsilon_0^2 \| \boldsymbol{\beta} - \boldsymbol{\beta}^* \|_{\boldsymbol{\Sigma}} \sqrt{\frac{2(d+t)}{n}} \tag{4.4}$$

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

Propositions 4.1 and 4.2 provide the key ingredients to prove Theorems 2.1 and 2.2. 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 4.3 and 4.4 below.

Recall that $\mathbb{P}^*$ and $\mathbb{E}^*$ denote, respectively, the probability measure and expectation (over $\mathcal{R}_n = \{e_i\}_{i=1}^n$) conditioning on $\mathcal{D}_n = \{(y_i, \boldsymbol{x}_i)\}_{i=1}^n$. For $i = 1, \ldots, n$, define

$$\zeta_i = I(\varepsilon_i \leqslant 0) - \tau \ \text{ and } \ z_i = \boldsymbol{\Sigma}^{-1/2} \boldsymbol{x}_i, \tag{4.5}$$

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

PROPOSITION 4.3 Assume Conditions 1 and 2 hold, and let $\boldsymbol{\xi}^\flat \in \partial Q_n^\flat(\boldsymbol{\beta}^*)$. For any $t > 0$, there exists some event $\mathcal{G}_1(t) = \mathcal{G}_1(t; \mathcal{D}_n)$ with $\mathbb{P}\{\mathcal{G}_1(t)\} \geqslant 1 - e^{-2t}$ such that, with $\mathbb{P}^*$-probability at least $1 - e^{-2t}$ conditioned on $\mathcal{G}_1(t)$,

$$\| \boldsymbol{\Sigma}^{-1/2} (\boldsymbol{\xi}^\flat - \mathbb{E}^* \boldsymbol{\xi}^\flat) \|_2 \leqslant 2\sqrt{\frac{d+t}{n}} \tag{4.6}$$

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

Similarly to Proposition 4.2, the following result establishes the RSC for the weighted quantile loss function.

PROPOSITION 4.4 Assume Conditions 1 and 2 hold. For any $t \geqslant 0$, there exists some event $\mathcal{G}_2(t) = \mathcal{G}_2(t; \mathcal{D}_n)$ such that $\mathbb{P}\{\mathcal{G}_2(t)\} \geqslant 1 - e^{-t}$, and with $\mathbb{P}^*$-probability at least $1 - e^{-t}/2$ conditioned on $\mathcal{G}_2(t)$,

$$\langle \boldsymbol{\xi}_{\boldsymbol{\beta}}^{\flat} - \boldsymbol{\xi}_{\boldsymbol{\beta}^*}^{\flat}, \boldsymbol{\beta} - \boldsymbol{\beta}^* \rangle \geqslant \frac{1}{8} \underline{f} \, \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma}}^2 - 8\upsilon_0^2 \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma}} \sqrt{\frac{2(d+t)}{n}} \tag{4.7}$$

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

Proofs of Propositions 4.1– 4.4 are placed in the Appendix.

4.2 *Proof of Theorem 2.1*

By the convexity of $\boldsymbol{\beta} \mapsto Q_n(\boldsymbol{\beta})$, $\hat{\boldsymbol{\beta}}$ satisfies the first-order condition that $\boldsymbol{\xi}_{\hat{\boldsymbol{\beta}}} = \mathbf{0}$ for some $\boldsymbol{\xi}_{\hat{\boldsymbol{\beta}}} \in \partial Q_n(\hat{\boldsymbol{\beta}})$. 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, $D(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) \geqslant 0$ for any subdifferentials $\boldsymbol{\xi}_{\boldsymbol{\beta}_1}$ and $\boldsymbol{\xi}_{\boldsymbol{\beta}_2}$. Taking $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = (\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}^*)$, we have

$$0 \leqslant \langle \boldsymbol{\xi}_{\hat{\boldsymbol{\beta}}} - \boldsymbol{\xi}_{\boldsymbol{\beta}^*}, \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \rangle = \langle -\boldsymbol{\xi}_{\boldsymbol{\beta}^*}, \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \rangle \leqslant \|\boldsymbol{\Sigma}^{-1/2} \boldsymbol{\xi}_{\boldsymbol{\beta}^*}\|_2 \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma}}, \tag{4.8}$$

for any $\boldsymbol{\xi}_{\boldsymbol{\beta}^*} \in \partial Q_n(\boldsymbol{\beta}^*)$. Starting with (4.8), we bound the left- and right-hand sides of (4.9) separately. To establish the lower bound, we use a localized argument (Sun *et al.*, 2019) and a new RSC property for the empirical quantile loss (Proposition 4.2).

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

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

By the above definition, $\eta = 1$ if $\hat{\boldsymbol{\beta}} \in \boldsymbol{\beta}^* + \mathbb{B}_{\boldsymbol{\Sigma}}(r)$ and $\eta < 1$ if $\hat{\boldsymbol{\beta}} \notin \boldsymbol{\beta}^* + \mathbb{B}_{\boldsymbol{\Sigma}}(r)$. In the latter case, we have $\widetilde{\boldsymbol{\beta}} \in \boldsymbol{\beta}^* + \partial\mathbb{B}_{\boldsymbol{\Sigma}}(r)$. Applying Lemma C.1 in Sun *et al.* (2019) with slight modifications yields the bound $D(\widetilde{\boldsymbol{\beta}}, \boldsymbol{\beta}^*) \leqslant \eta D(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}^*)$, leading to

$$\langle \boldsymbol{\xi}_{\widetilde{\boldsymbol{\beta}}} - \boldsymbol{\xi}_{\boldsymbol{\beta}^*}, \widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \rangle \leqslant \eta \langle \boldsymbol{\xi}_{\hat{\boldsymbol{\beta}}} - \boldsymbol{\xi}_{\boldsymbol{\beta}^*}, \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \rangle, \tag{4.9}$$

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 $\boldsymbol{\xi}_{\hat{\boldsymbol{\beta}}} = \mathbf{0}$ and Cauchy–Schwarz inequality, implies

$$\langle \boldsymbol{\xi}_{\widetilde{\boldsymbol{\beta}}} - \boldsymbol{\xi}_{\boldsymbol{\beta}^*}, \widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \rangle \leqslant \eta \langle -\boldsymbol{\xi}_{\boldsymbol{\beta}^*}, \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \rangle \leqslant \|\boldsymbol{\Sigma}^{-1/2} \boldsymbol{\xi}_{\boldsymbol{\beta}^*}\|_2 \|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma}}. \tag{4.10}$$

Note that (4.10) is a localized version of (4.8) 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 4.2 that

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

with probability at least $1 - e^{-t}/2$. Combining this with (4.9) and (4.10), and taking $x = t > 0$ in Proposition 4.1, we obtain

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

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

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

with probability at least $1 - 2e^{-t}$ as long as $n \geq CL_0^2 \underline{f}^{-4}(d+t)$ 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 $\hat{\boldsymbol{\beta}} = \widetilde{\boldsymbol{\beta}} \in \boldsymbol{\beta}^* + \mathbb{B}_{\boldsymbol{\Sigma}}(r)$. Otherwise if $\hat{\boldsymbol{\beta}} \notin \boldsymbol{\beta}^* + \mathbb{B}_{\boldsymbol{\Sigma}}(r)$, 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.

4.3 *Proof of Theorem 2.2*

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

$$\nabla Q_n(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i \big\{ I(y_i \leqslant \langle \boldsymbol{x}_i, \boldsymbol{\beta} \rangle) - \tau \big\}, \quad \boldsymbol{\beta} \in \mathbb{R}^d. \tag{4.11}$$

Recall from Condition 2 that the conditional distribution of ε given $\boldsymbol{x}$ is continuous. Lemma A.1 in Ruppert & Carroll (1980) states that with probability one, there is no vector $\boldsymbol{\delta} \in \mathbb{R}^d$ and $1 \leqslant i \leqslant n$ such that $\varepsilon_i = \langle \boldsymbol{x}_i, \boldsymbol{\delta} \rangle$. It follows 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

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

Next, define the vector-valued random process

$$\Delta(\boldsymbol{\beta}) = \mathbf{S}^{-1/2}\big\{\nabla Q_n(\boldsymbol{\beta}) - \nabla Q_n(\boldsymbol{\beta}^*)\big\} - \mathbf{S}^{1/2}(\boldsymbol{\beta} - \boldsymbol{\beta}^*), \tag{4.12}$$

where $\mathbf{S} = \nabla^2 \mathbb{E}Q_n(\boldsymbol{\beta}^*) = \mathbb{E}\{f_{\varepsilon|\boldsymbol{x}}(0)\boldsymbol{x}\boldsymbol{x}^\intercal\}$. 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

$$\begin{aligned}\mathbb{E}\Delta(\boldsymbol{\beta}) &= \mathbf{S}^{-1/2}\left\langle \int_0^1 \nabla^2 \mathbb{E}Q_n(\boldsymbol{\beta}_t^*)\mathrm{d}t, \boldsymbol{\beta}-\boldsymbol{\beta}^*\right\rangle - \mathbf{S}^{1/2}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)\\ &= \left\langle \mathbf{S}^{-1/2}\int_0^1 \nabla^2 \mathbb{E}Q_n(\boldsymbol{\beta}_t^*)\mathrm{d}t\,\mathbf{S}^{-1/2} - \mathbf{i}_d, \mathbf{S}^{1/2}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)\right\rangle, \end{aligned} \tag{4.13}$$

where $\boldsymbol{\beta}_t^* = (1-t)\boldsymbol{\beta}^* + t\boldsymbol{\beta}$ and $\nabla^2\mathbb{E}Q_n(\boldsymbol{\beta}_t^*) = \mathbb{E}\{f_{\varepsilon|\boldsymbol{x}}(t\langle \boldsymbol{x}, \boldsymbol{\beta}-\boldsymbol{\beta}^*\rangle)\boldsymbol{x}\boldsymbol{x}^\intercal\}$. For $r>0$, define the local elliptic neighborhood of $\boldsymbol{\beta}^*$ as $\Theta_{\boldsymbol{\Sigma}}(r) := \{\boldsymbol{\beta}\in\mathbb{R}^d : \|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma}} \leqslant r\}$. By Conditions 1 and 2, $\boldsymbol{\Sigma}$ is positive definite and $\underline{f} \leqslant f_{\varepsilon|\boldsymbol{x}}(0) \leqslant \bar{f}$, 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

$$\begin{aligned}&\left\|\mathbf{S}^{-1/2}\nabla^2\mathbb{E}Q_n(\boldsymbol{\beta}_t^*)\mathbf{S}^{-1/2} - \mathbf{i}_d\right\|_2 = \left\|\mathbf{S}^{-1/2}\mathbb{E}\left[\left\{f_{\varepsilon|\boldsymbol{x}}(t\langle\boldsymbol{x},\boldsymbol{\delta}\rangle) - f_{\varepsilon|\boldsymbol{x}}(0)\right\}\boldsymbol{x}\boldsymbol{x}^\intercal\right]\mathbf{S}^{-1/2}\right\|_2\\ &\leqslant L_0 t\cdot \sup_{\boldsymbol{u}\in\mathbb{B}^d(1)} \mathbb{E}\left\{\langle \mathbf{S}^{-1/2}\boldsymbol{x},\boldsymbol{u}\rangle^2|\langle\boldsymbol{x},\boldsymbol{\delta}\rangle|\right\} \leqslant \underline{f}^{-1}L_0 t\cdot\left(\sup_{\boldsymbol{u}\in\mathbb{B}^d(1)}\mathbb{E}|\langle\boldsymbol{\Sigma}^{-1/2}\boldsymbol{x},\boldsymbol{u}\rangle|^3\right)^{2/3}\left(\mathbb{E}|\langle\boldsymbol{x},\boldsymbol{\delta}\rangle|^3\right)^{1/3}\\ &\leqslant L_0\underline{f}^{-1}m_3 rt,\end{aligned}$$

where $m_k := \sup_{\boldsymbol{u}\in\mathbb{B}^d(1)}\mathbb{E}|\langle\boldsymbol{\Sigma}^{-1/2}\boldsymbol{x},\boldsymbol{u}\rangle|^k$ (for $k\geqslant 1$) depends only on υ_0 and k. Combining this with (4.13), we obtain

$$\sup_{\boldsymbol{\beta}\in\Theta_{\boldsymbol{\Sigma}}(r)}\|\mathbb{E}\Delta(\boldsymbol{\beta})\|_2 \leqslant \frac{1}{2}L_0\underline{f}^{-1}\bar{f}^{1/2}m_3 r^2. \tag{4.14}$$

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

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

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

$$\begin{aligned}\sup_{\boldsymbol{\beta}\in\Theta_{\boldsymbol{\Sigma}}(r)}\|\Delta(\boldsymbol{\beta}) - \mathbb{E}\Delta(\boldsymbol{\beta})\|_2 &\leqslant \underline{f}^{-1/2}\sup_{\boldsymbol{\beta}\in\Theta_{\boldsymbol{\Sigma}}(r)}\|\boldsymbol{\Sigma}^{-1/2}\{R_n(\boldsymbol{\beta}) - R_n(\boldsymbol{\beta}^*)\}\|_2\\ &= \underline{f}^{-1/2}\sup_{\boldsymbol{v}\in\mathbb{B}^d(r)}\|\boldsymbol{\Sigma}^{-1/2}\{R_n(\boldsymbol{\beta}^*+\boldsymbol{\Sigma}^{-1/2}\boldsymbol{v}) - R_n(\boldsymbol{\beta}^*)\}\|_2\\ &= \underline{f}^{-1/2}r^{-1}\sup_{\boldsymbol{u},\boldsymbol{v}\in\mathbb{B}^d(r)}\underbrace{\langle\boldsymbol{\Sigma}^{-1/2}\{R_n(\boldsymbol{\beta}^*+\boldsymbol{\Sigma}^{-1/2}\boldsymbol{v}) - R_n(\boldsymbol{\beta}^*)\},\boldsymbol{u}\rangle}_{n^{-1/2}\Delta_0(\boldsymbol{u},\boldsymbol{v})},\end{aligned} \tag{4.15}$$

where $\Delta_0(\boldsymbol{u},\boldsymbol{v}) = n^{-1/2}\sum_{i=1}^n(1-\mathbb{E})\langle\boldsymbol{z}_i,\boldsymbol{u}\rangle\{I(\varepsilon_i \leqslant \langle\boldsymbol{z}_i,\boldsymbol{v}\rangle) - I(\varepsilon_i\leq 0)\}$. To bound $\sup_{\boldsymbol{u},\boldsymbol{v}\in\mathbb{B}^d(r)}\Delta_0(\boldsymbol{u},\boldsymbol{v})$, 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 & Wellner, 1996). Consider the following two classes of real-valued functions on $\mathbb{R} \times \mathbb{R}^d$:

$$\mathcal{F}_1 = \{(z_0, \boldsymbol{z}) \mapsto \langle \boldsymbol{z}, \boldsymbol{u} \rangle : \boldsymbol{u} \in \mathbb{B}^d(r)\} \text{ and } \mathcal{F}_2 = \{(z_0, \boldsymbol{z}) \mapsto I(\langle \boldsymbol{z}, \boldsymbol{v} \rangle - z_0 \geqslant 0) : \boldsymbol{v} \in \mathbb{B}^d(r)\}. \tag{4.16}$$

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

$$\sup_{\boldsymbol{v}, \boldsymbol{u} \in \mathbb{B}^d(r)} \Delta_0(\boldsymbol{u}, \boldsymbol{v}) = \sup_{f \in \mathcal{F}} \underbrace{\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{f(\bar{z}_i) - \mathbb{E} f(\bar{z}_i)\}}_{\mathbb{G}_n f}, \tag{4.17}$$

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

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

where $c_0 > 0$ depends only on υ_0, and $\| \cdot \|_{\psi_q}$ $(1 \leqslant q \leqslant 2)$ 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 $x \geqslant 0$,

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

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

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

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

$$\sup_{f \in \mathcal{F}} \mathbb{G}_n f \leqslant \frac{3}{2} \mathbb{E} \left(\sup_{f \in \mathcal{F}} \mathbb{G}_n f \right) + \max \left\{ (3M_0 m_3)^{1/2} r^{3/2} t^{1/2}, 2c_1 r t \sqrt{\frac{d \log n}{n}} \right\}. \tag{4.18}$$

To bound $\mathbb{E}(\sup_{f \in \mathcal{F}} \mathbb{G}_n f)$, the key is to control the covering numbers $N(\mathcal{F}, L_2(Q), \epsilon \|F\|_{Q,2})$ for all finitely supported probability measures Q on $\mathbb{R} \times \mathbb{R}^d$ and $0 < \epsilon < 1$, where $F(\bar{z}) = r\|\boldsymbol{z}\|_2$ is a measurable envelope of $\mathcal{F}$. Respectively, for the function classes $\mathcal{F}_1$ and $\mathcal{F}_2$ that have envelopes $F_1(\bar{z}) = r\|\boldsymbol{z}\|_2$ and

$F_2(\bar{z}) = 1$, using Theorem B in Dudley (1979) and Theorem 2.6.7 in van der Vaart & Wellner (1996) we have

$$\sup_Q N(\mathcal{F}_1, L_2(Q), \epsilon \|F_1\|_{Q,2}) \leqslant (A_1/\epsilon)^{2(d+2)} \text{ and } \sup_Q N(\mathcal{F}_2, L_2(Q), \epsilon) \leqslant (A_1/\epsilon)^{2(d+2)}$$

for some $A_1 > e$, where the suprema are taken over all finitely discrete probability measures Q on $\mathbb{R} \times \mathbb{R}^d$. Combining the above bounds with Corollary A.1 in the supplement of Chernozhukov *et al.* (2014) shows that

$$\begin{aligned}&\sup_Q N(\mathcal{F}, L_2(Q), \epsilon \|F\|_{Q,2}) \\ &\leqslant \sup_Q N(\mathcal{F}_1, L_2(Q), 2^{-1/2}\epsilon \|F_1\|_{Q,2}) \cdot \sup_Q N(\mathcal{F}_2, L_2(Q), 2^{-1/2}\epsilon) \leqslant (A_2/\epsilon)^{4(d+2)},\end{aligned}$$

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

$$\mathbb{E}\left(\sup_{f\in\mathcal{F}} \mathbb{G}_n f\right) \lesssim \sqrt{M_0 m_3 r^3 d \log\left(A_2^2 d/(M_0 m_3 r)\right)} + r M_n \frac{d}{n^{1/2}} \log\left(A_2^2 d/(M_0 m_3 r)\right), \tag{4.19}$$

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

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

Given $\epsilon \in (0,1)$, there exits a finite subset $\mathcal{N}_\epsilon \subseteq \mathbb{S}^{d-1}$ with $|\mathcal{N}_\epsilon| \leqslant (1 + 2/\epsilon)^d$ such that $\max_{1\leqslant i\leqslant n} \|\boldsymbol{z}_i\|_2 \leqslant (1-\epsilon)^{-1} \max_{1\leqslant i\leqslant n} \max_{\boldsymbol{u}\in\mathcal{N}_\epsilon} \langle \boldsymbol{u}, \boldsymbol{w}_i\rangle$. For every $i \in [n]$ and $\boldsymbol{u} \in \mathcal{N}_\epsilon$, Condition 1 indicates that $\mathbb{P}(|\langle \boldsymbol{u}, \boldsymbol{w}_i\rangle| \geqslant \upsilon_0 u) \leqslant 2e^{-u^2/2}$ for any $u \in \mathbb{R}$. Taking the union bound over $i \in [n]$ and $\boldsymbol{u} \in \mathcal{N}_\epsilon$, and setting $u = \sqrt{2v + 2\log(2n) + 2d\log(1 + 2/\epsilon)}$ $(v > 0)$, we obtain that with probability at least $1 - 2n(1+2/\epsilon)^d e^{-u^2/2} = 1 - e^{-v}$, $\max_{1\leqslant i\leqslant n} \|\boldsymbol{z}_i\|_2 \leqslant (1-\epsilon)^{-1}\upsilon_0\sqrt{2v + 2\log(2n) + 2d\log(1+2/\epsilon)}$. Minimizing this upper bound with respect to $\epsilon \in (0,1)$, we conclude that

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

Taking $A = 2\upsilon_0^2\{3.7d + \log(2n)\}$ and $a = 2\upsilon_0^2$ in the earlier analysis yields the bound $M_n^2 = \mathbb{E}(\max_{1\leqslant i\leqslant n}\|z_i\|_2^2) \leqslant 2\upsilon_0^2\{3.7d + \log(2en)\}$. Plugging this into (4.19) gives

$$\mathbb{E}\left(\sup_{f\in\mathcal{F}}\mathbb{G}_n f\right) \lesssim \sqrt{M_0 m_3 r^3 d\log\left(A_2^2 d/(M_0 m_3 r)\right)} + r(d+\log n)^{1/2}\frac{d}{n^{1/2}}\log\left(A_2^2 d/(M_0 m_3 r)\right). \quad (4.20)$$

Together, (4.15), (4.17), (4.18) and (4.20) imply that with probability at least $1 - 2e^{-t}$,

$$\begin{aligned}&\sup_{\boldsymbol{\beta}\in\Theta_{\boldsymbol{\Sigma}}(r)}\|\Delta(\boldsymbol{\beta}) - \mathbb{E}\Delta(\boldsymbol{\beta})\|_2\\ &\leqslant C_1\left\{\sqrt{\frac{rt}{n}} + \sqrt{\log(C_2 d/r)\frac{rd}{n}} + (d+\log n)^{1/2}\log(C_2 d/r)\frac{d}{n} + (d\log n)^{1/2}\frac{t}{n}\right\}. \quad (4.21)\end{aligned}$$

Thus far, we have established a high probability bound on the ℓ_2-norm of $\Delta(\boldsymbol{\beta}) = \mathbf{S}^{-1/2}\{\nabla Q_n(\boldsymbol{\beta}) - \nabla Q_n(\boldsymbol{\beta}^*)\} - \mathbf{S}^{1/2}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)$ uniformly over $\boldsymbol{\beta}\in\Theta_{\boldsymbol{\Sigma}}(r)$, a local neighborhood of $\boldsymbol{\beta}^*$, for any prespecified $r > 0$. By Theorem 2.1, we have $\hat{\boldsymbol{\beta}}\in\Theta_{\boldsymbol{\Sigma}}(r_t)$ with probability at least $1 - 2e^{-t}$ as long as $n \geqslant CL_0^2 f^{-4}(d+t)$, where $r_t = C_3\sqrt{(d+t)/n}$. Setting $r = r_t$ in (4.14) and (4.21), we find that with probability at least $1 - 2e^{-t}$,

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

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

4.4 *Proof of Theorem 2.3*

Let $\boldsymbol{\lambda}\in\mathbb{R}^d$ 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 $\zeta_i = I(\varepsilon_i \leqslant 0) - \tau$ and $\gamma_i = -\langle\mathbf{S}^{-1}\boldsymbol{\lambda}, \boldsymbol{x}_i\rangle$. Moreover, write $\delta_n = (d+\log n)^{1/4}(d\log n)^{1/2}n^{-1/4} + (d+\log n)^{1/2}d\log(n)\,n^{-1/2}$. Applying Theorem 2.2 with $t = \log n$ yields that, under the scaling $n \gtrsim d + \log n$,

$$\begin{aligned}&\left|n^{1/2}\langle\boldsymbol{\lambda}, \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\rangle - S_n\right|\\ &= n^{1/2}\left|\left\langle\mathbf{S}^{-1/2}\boldsymbol{\lambda}, \mathbf{S}^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) + \mathbf{S}^{-1/2}\frac{1}{n}\sum_{i=1}^n\{I(\varepsilon_i \leqslant 0) - \tau\}\boldsymbol{x}_i\right\rangle\right| \leqslant c_1\|\mathbf{S}^{-1/2}\boldsymbol{\lambda}\|_2\,\delta_n \quad (4.22)\end{aligned}$$

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

$$\begin{aligned}&\sup_{x\in\mathbb{R}}\big|\mathbb{P}\big\{S_n \leqslant \mathrm{var}(S_n)^{1/2}x\big\} - \Phi(x)\big| \\ &\leqslant \frac{\mathbb{E}|\{I(\varepsilon \leqslant 0)-\tau\}\langle \mathbf{S}^{-1}\boldsymbol{\lambda}, \boldsymbol{x}\rangle|^3}{2n^{1/2}\sigma_\tau^3} \leqslant \frac{1-2(\tau-\tau^2)}{2(\tau-\tau^2)^{1/2}}\frac{m_3}{n^{1/2}} = c_2 n^{-1/2}. \end{aligned} \tag{4.23}$$

Moreover, for any $a \leqslant b$, $\Phi(b/\sigma_\tau) - \Phi(a/\sigma_\tau) \leqslant (2\pi)^{-1/2}(b-a)/\sigma_\tau$. Combining this with (4.22) and (4.23), for any $x \in \mathbb{R}$, we obtain

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

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

4.5 *Proof of Theorem 2.4*

Without loss of generality, we assume $t > 0$ is such that $2e^{-t} \leqslant 1$ throughout the proof. By the convexity of $\boldsymbol{\beta} \mapsto Q_n^\flat(\boldsymbol{\beta})$, $\hat{\boldsymbol{\beta}}^\flat$ satisfies the first-order condition that $\boldsymbol{\xi}^\flat_{\hat{\boldsymbol{\beta}}^\flat} = \mathbf{0}$ for some $\boldsymbol{\xi}^\flat_{\hat{\boldsymbol{\beta}}^\flat} \in \partial Q_n^\flat(\hat{\boldsymbol{\beta}}^\flat)$. Again, we follow the same localized analysis as in the proof of Theorem 2.1. For some $0 < r \leqslant \underline{f}/(6L_0\upsilon_0^2)$ to be determined, if $\hat{\boldsymbol{\beta}}^\flat \notin \boldsymbol{\beta}^* + \mathbb{B}_{\boldsymbol{\Sigma}}(r)$, there exists $\eta \in (0,1)$ such that $\widetilde{\boldsymbol{\beta}} := \boldsymbol{\beta}^* + \eta(\hat{\boldsymbol{\beta}}^\flat - \boldsymbol{\beta}^*) \in \boldsymbol{\beta}^* + \partial\mathbb{B}_{\boldsymbol{\Sigma}}(r)$; otherwise if $\hat{\boldsymbol{\beta}}^\flat \in \boldsymbol{\beta}^* + \mathbb{B}_{\boldsymbol{\Sigma}}(r)$, we take $\eta = 1$ so that $\widetilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^\flat$.

Similar to (4.9) and (4.10), we have that for any $\boldsymbol{\xi}^\flat_{\boldsymbol{\beta}^*} \in \partial Q_n^\flat(\boldsymbol{\beta}^*)$ and $\boldsymbol{\xi}^\flat_{\widetilde{\boldsymbol{\beta}}} \in \partial Q_n^\flat(\widetilde{\boldsymbol{\beta}})$,

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

For the right-hand side, Proposition 4.3 implies that there exists some event $\mathcal{G}_1(t)$ with $\mathbb{P}\{\mathcal{G}_1(t)\} \geqslant 1 - e^{-2t}$ such that, conditioned on $\mathcal{G}_1(t)$,

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

with $\mathbb{P}^*$-probability at least $1 - e^{-2t}$ as long as $n \gtrsim d + t$. On the other hand, since $\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma}} \leqslant r$, by Proposition 4.4, there exists some event $\mathcal{G}_2(t) = \mathcal{G}_2(t; \mathcal{D}_n)$ with $\mathbb{P}\{\mathcal{G}_2(t)\} \geqslant 1 - e^{-t}$ such that,

conditioned on $\mathcal{G}_2(t)$,

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

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

$$\|\widetilde{\boldsymbol{\delta}}\|_{\boldsymbol{\Sigma}} \leq 8\underline{f}^{-1}\big(2^{1/2} + 8\upsilon_0^2\big)\sqrt{\frac{2(d+t)}{n}} + 8\underline{f}^{-1}\|\boldsymbol{\Sigma}^{-1/2}\mathbb{E}^*\boldsymbol{\xi}^{\flat}_{\boldsymbol{\beta}^*}\|_2 \tag{4.24}$$

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

For $\|\boldsymbol{\Sigma}^{-1/2}\mathbb{E}^*\boldsymbol{\xi}^{\flat}_{\boldsymbol{\beta}^*}\|_2$, it follows from (4.3) and Proposition 4.1 that

$$\|\boldsymbol{\Sigma}^{-1/2}\mathbb{E}^*\boldsymbol{\xi}^{\flat}_{\boldsymbol{\beta}^*}\|_2 < 3\upsilon_0\sqrt{\frac{2(d+t)}{n}} \tag{4.25}$$

with probability at least $1 - e^{-2t}$. Let $\mathcal{G}_3(t)$ be the event that (4.25) holds so that $\mathbb{P}\{\mathcal{G}_3(t)\} \geqslant 1 - e^{-2t}$.

Combining (4.24) and (4.25), we conclude that conditioned on $\mathcal{G}_1(t) \cap \mathcal{G}_2(t) \cap \mathcal{G}_3(t)$, $\|\widetilde{\boldsymbol{\delta}}\|_{\boldsymbol{\Sigma}} < r := C_4\underline{f}^{-1}\sqrt{(d+t)/n}$ with $\mathbb{P}^*$-probability at least $1 - e^{-t}$ as long as $n \geqslant C_5 L_0^2 \underline{f}^{-4}(d+t)$, and $\mathbb{P}\{\mathcal{G}_1(t) \cap \mathcal{G}_2(t) \cap \mathcal{G}_3(t)\} \geqslant 1 - 2e^{-t}$, where the constants $C_4, C_5 > 0$ depend only on υ_0. This enforces $\widetilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^{\flat}$. Finally, taking $\mathcal{E}(t) = \mathcal{G}_1(t) \cap \mathcal{G}_2(t) \cap \mathcal{G}_3(t)$ establishes the claim.

4.6 *Proof of Theorem 2.5*

Following the proof of Theorem 2.2, we treat $\nabla Q^{\flat}_n(\boldsymbol{\beta}) := (1/n)\sum_{i=1}^n w_i \boldsymbol{x}_i\{I(y_i \leqslant \langle \boldsymbol{x}_i, \boldsymbol{\beta}\rangle) - \tau\}$ as the gradient of $Q^{\flat}_n(\boldsymbol{\beta})$. Under this notation, define the vector-valued random process

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

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

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

and $\mathbb{E}^*\Delta^{\flat}(\boldsymbol{\beta}) = \Delta(\boldsymbol{\beta})$ with $\Delta(\boldsymbol{\beta})$ defined in (4.12). By the triangle inequality, for any $r > 0$ we have

$$\sup_{\boldsymbol{\beta} \in \Theta_{\boldsymbol{\Sigma}}(r)} \|\Delta^{\flat}(\boldsymbol{\beta})\|_2 \leqslant \sup_{\boldsymbol{\beta} \in \Theta_{\boldsymbol{\Sigma}}(r)} \|\Delta^{\flat}(\boldsymbol{\beta}) - \mathbb{E}^*\Delta^{\flat}(\boldsymbol{\beta})\|_2 + \sup_{\boldsymbol{\beta} \in \Theta_{\boldsymbol{\Sigma}}(r)} \|\Delta(\boldsymbol{\beta})\|_2, \tag{4.26}$$

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

The last term $\sup_{\boldsymbol{\beta} \in \Theta_{\boldsymbol{\Sigma}}(r)} \|\Delta(\boldsymbol{\beta})\|_2$ in (4.26), which only depends on the data $\mathcal{D}_n = \{(y_i, \boldsymbol{x}_i)\}_{i=1}^n$, has been dealt with in the proof of Theorem 2.2. 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 $\mathcal{D}_n$. As before, we use a change of variable $\boldsymbol{v} = \boldsymbol{\Sigma}^{1/2}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)$ and obtain

$$\begin{aligned}
&\sup_{\boldsymbol{\beta}\in\Theta_{\boldsymbol{\Sigma}}(r)} \|\Delta^\flat(\boldsymbol{\beta}) - \mathbb{E}^*\Delta^\flat(\boldsymbol{\beta})\|_2 = \sup_{\boldsymbol{\beta}\in\Theta_{\boldsymbol{\Sigma}}(r)} \|\mathbf{S}^{-1/2}\{R_n^\flat(\boldsymbol{\beta}) - R_n^\flat(\boldsymbol{\beta}^*)\}\|_2 \\
&\leqslant \underline{f}^{-1/2} \sup_{\boldsymbol{\beta}\in\Theta_{\boldsymbol{\Sigma}}(r),\boldsymbol{u}\in\mathbb{B}^d(1)} \langle R_n^\flat(\boldsymbol{\beta}) - R_n^\flat(\boldsymbol{\beta}^*), \boldsymbol{\Sigma}^{-1/2}\boldsymbol{u}\rangle \\
&= \underline{f}^{-1/2} r^{-1} \sup_{\boldsymbol{u},\boldsymbol{v}\in\mathbb{B}^d(r)} \underbrace{\langle \boldsymbol{\Sigma}^{-1/2}\{R_n^\flat(\boldsymbol{\beta}^* + \boldsymbol{\Sigma}^{-1/2}\boldsymbol{v}) - R_n^\flat(\boldsymbol{\beta}^*)\}, \boldsymbol{v}\rangle}_{n^{-1/2}\Delta_0^\flat(\boldsymbol{u},\boldsymbol{v})},
\end{aligned} \tag{4.27}$$

where $\Delta_0^\flat(\boldsymbol{u},\boldsymbol{v}) = n^{-1/2}\sum_{i=1}^n e_i\langle \boldsymbol{z}_i, \boldsymbol{u}\rangle\{I(\varepsilon_i \leqslant \langle \boldsymbol{z}_i, \boldsymbol{v}\rangle) - I(\varepsilon_i \leqslant 0)\}$. Let $\mathcal{F}_1$ and $\mathcal{F}_2$ be the function classes defined in (4.16), and let $\mathcal{F} = \mathcal{F}_1 \cdot (\mathcal{F}_2 - f_0)$ be the pointwise product between $\mathcal{F}_1$ and $\mathcal{F}_2 - f_0$ with $f_0 : (z_0, \boldsymbol{z}) \mapsto I(z_0 \leqslant 0)$. With this notation, we have $\sup_{\boldsymbol{u},\boldsymbol{v}\in\mathbb{B}^d(r)} \Delta_0^\flat(\boldsymbol{u},\boldsymbol{v}) = \sup_{f\in\mathcal{F}} n^{-1/2}\sum_{i=1}^n e_i f(\bar{\boldsymbol{z}}_i)$. Recall that $\mathbb{E}^*$ denotes the conditional expectation given $\mathcal{D}_n$. By Theorem 13 in Boucheron *et al.* (2005) and the bound $\sup_{1\leqslant i\leqslant n, f\in\mathcal{F}} f(\bar{\boldsymbol{z}}_i) \leqslant r\max_{1\leqslant i\leqslant n}\|\boldsymbol{z}_i\|_2$, we obtain that, with $Z := \mathbb{E}^*\{\sup_{f\in\mathcal{F}} |(1/n)\sum_{i=1}^n e_i f(\bar{\boldsymbol{z}}_i)|\}$ denoting the conditional Rademacher average,

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

where $\kappa = \sqrt{e}/(2\sqrt{e} - 2) < 1.271$ and $M_{n,k} := (\mathbb{E}\max_{1\leqslant i\leqslant n}\|\boldsymbol{z}_i\|_2^{2k})^{1/(2k)}$. By (4.27), Markov's inequality and the bound $Z \leqslant (Z - \mathbb{E}Z)_+ + \mathbb{E}Z$, we obtain that

$$\sup_{\boldsymbol{\beta}\in\Theta_{\boldsymbol{\Sigma}}(r)} \|\Delta^\flat(\boldsymbol{\beta}) - \mathbb{E}^*\Delta^\flat(\boldsymbol{\beta})\|_2 = O_{\mathbb{P}^*}\big(r^{-1}Z\big) \text{ and } Z = O_{\mathbb{P}}\big(\mathbb{E}Z + rM_{n,1}/n\big). \tag{4.28}$$

For $\mathbb{E}Z$, by a similar argument to (4.20) and (4.21), we get

$$\mathbb{E}Z \lesssim r^{3/2}\sqrt{\log(C_2 d/r)\frac{d}{n}} + r(d + \log n)^{1/2}\log(C_2 d/r)\frac{d}{n}. \tag{4.29}$$

With the above preparations, we are ready to prove the claim. Together, Theorems 2.1– 2.4 imply that under the scaling $n \gtrsim d + \log n$, there exists some event $\mathcal{E}_n$, satisfying $\mathbb{P}(\mathcal{E}_n) \geqslant 1 - 4n^{-1}$, on which $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma}} \leqslant r_n = C_3\sqrt{(d + \log n)/n}$ and

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

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

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

By (4.26), (4.28), (4.29) and (4.21), $\chi_{2n} = \chi_{2n}(\mathcal{D}_n) := \mathbb{E}^*\{\sup_{\boldsymbol{\beta}\in\Theta_{\boldsymbol{\Sigma}}(r_n^\flat)} \|\Delta^\flat(\boldsymbol{\beta})\|_2\}$ satisfies $\chi_{2n} = O_{\mathbb{P}}(\Delta_{n,d})$. Let $\boldsymbol{r}_n^\flat = \mathbf{S}^{1/2}(\hat{\boldsymbol{\beta}}^\flat - \hat{\boldsymbol{\beta}}) - \mathbf{S}^{-1/2}(1/n)\sum_{i=1}^n e_i\boldsymbol{x}_i\{\tau - I(\varepsilon_i \leqslant 0)\}$. Then, with $\mathbb{P}^*$-probability at least $1-n^{-1}$ conditioned on $\mathcal{E}_n$, $\|\boldsymbol{r}_n^\flat\|_2 \leqslant \chi_{1n} + \sup_{\boldsymbol{\beta}\in\Theta_{\boldsymbol{\Sigma}}(r)} \|\Delta^\flat(\boldsymbol{\beta})\|_2$ with $\sup_{\boldsymbol{\beta}\in\Theta_{\boldsymbol{\Sigma}}(r)} \|\Delta^\flat(\boldsymbol{\beta})\|_2 = O_{\mathbb{P}^*}(\chi_{2n})$ and $\chi_{1n} + \chi_{2n} = O_{\mathbb{P}}(\Delta_{n,d})$. This establishes the claim (2.14).

4.7 *Proof of Theorem 2.6*

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

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

To begin with, it follows from Theorem 2.2 that under the scaling $n \gtrsim d + \log n$, there exists a sequence of events $\{\mathcal{E}_n\}$ with $\mathbb{P}(\mathcal{E}_n) \geqslant 1 - 4n^{-1}$ such that, $|n^{1/2}\langle \boldsymbol{\lambda}, \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\rangle - S_n| \leqslant c_1\|\mathbf{S}^{-1/2}\boldsymbol{\lambda}\|_2\delta_{n,d}$ on $\mathcal{E}_n$, where $\delta_{n,d} := (d+\log n)^{1/4}(d\log n)^{1/2}n^{-1/4} + (d+\log n)^{1/2}d\log(n)n^{-1/2}$. By Theorems 2.4 and 2.5, we further have $|n^{1/2}\langle \boldsymbol{\lambda}, \hat{\boldsymbol{\beta}}^\flat - \hat{\boldsymbol{\beta}}\rangle - S_n^\flat| \leqslant \|\mathbf{S}^{-1/2}\boldsymbol{\lambda}\|_2\|n^{1/2}\boldsymbol{r}_n^\flat\|_2$ with $\mathbb{P}^*$-probability at least $1-n^{-1}$ conditioned on $\mathcal{E}_n$. 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 (4.28) and (4.29), there exits some event $\mathcal{G}_n$ with $\mathbb{P}(\mathcal{G}_n^{\mathrm{c}}) \lesssim (\delta_{n,d}/\delta_2)^2$ such that, conditioned on $\mathcal{E}_n \cap \mathcal{G}_n$,

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

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

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

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

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

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

$$\sup_{x\in\mathbb{R}}\left|\mathbb{P}\left\{S_n \leqslant \mathrm{var}(S_n)^{1/2}x\right\} - \Phi(x)\right| \leqslant \frac{m_3}{2\sqrt{\tau(1-\tau)n}}. \tag{4.30}$$

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

$$\sup_{x\in\mathbb{R}}\left|\mathbb{P}^*\left\{S_n^\flat \leqslant \mathrm{var}^*(S_n^\flat)^{1/2}x\right\} - \Phi(x)\right| \leqslant \frac{(1/n)\sum_{i=1}^n|\zeta_i\gamma_i|^3}{2\sqrt{n}\,\{\mathrm{var}^*(S_n^\flat)\}^{3/2}}, \tag{4.31}$$

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 $\boldsymbol{u} = \boldsymbol{\Sigma}^{1/2}\mathbf{S}^{-1}\boldsymbol{\lambda}/\|\mathbf{S}^{-1}\boldsymbol{\lambda}\|_{\boldsymbol{\Sigma}} \in \mathbb{S}^{d-1}$ 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

$$\left|\mathrm{var}^*(S_n^\flat)/\mathrm{var}(S_n) - 1\right| = \frac{1}{\tau(1-\tau)}\left|\frac{1}{n}\sum_{i=1}^n\zeta_i^2\langle\boldsymbol{u},\boldsymbol{z}_i\rangle^2 - \tau(1-\tau)\right| \tag{4.32}$$

and

$$\frac{1}{n}\sum_{i=1}^n|\gamma_i\zeta_i|^3 \leqslant \max_{1\leqslant i\leqslant n}|\gamma_i\zeta_i|\cdot\frac{1}{n}\sum_{i=1}^n\zeta_i^2\langle\mathbf{S}^{-1}\boldsymbol{\lambda},\boldsymbol{x}_i\rangle^2 \leqslant \max_{1\leqslant i\leqslant n}|\gamma_i\zeta_i|\cdot\|\mathbf{S}^{-1}\boldsymbol{\lambda}\|_{\boldsymbol{\Sigma}}^2\cdot\frac{1}{n}\sum_{i=1}^n\zeta_i^2\langle\boldsymbol{u},\boldsymbol{z}_i\rangle^2. \tag{4.33}$$

For independent zero-mean sub-Gaussian random variables $\gamma_i\zeta_i$, it can be shown that with probability at least $1-e^{-x}$, $\max_{1\leqslant i\leqslant n}|\gamma_i\zeta_i| \lesssim \|\mathbf{S}^{-1}\boldsymbol{\lambda}\|_{\boldsymbol{\Sigma}}\sqrt{\log(n)+x}$. Furthermore, following the proof of Proposition 4.3, it can be similarly shown that

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

with probability at least $1-2e^{-x}$. Putting together the pieces, it follows from (4.32) that there exists an event $\mathcal{E}_n'$, satisfying $\mathbb{P}(\mathcal{E}_n') \geqslant 1-n^{-1}$, on which $\max_{1\leqslant i\leqslant n}|\gamma_i\zeta_i| \lesssim \|\mathbf{S}^{-1}\boldsymbol{\lambda}\|_{\boldsymbol{\Sigma}}(\log n)^{1/2}$,

$$\frac{1}{n}\sum_{i=1}^n|\gamma_i\zeta_i|^3 \lesssim \|\mathbf{S}^{-1}\boldsymbol{\lambda}\|_{\boldsymbol{\Sigma}}^3(\log n)^{1/2} \text{ and } \left|\mathrm{var}^*(S_n^\flat)/\mathrm{var}(S_n) - 1\right| \lesssim \sqrt{\frac{\log n}{n}} \tag{4.34}$$

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 & Zhilova, 2015):

$$\sup_{x\in\mathbb{R}}\left|\Phi(x/\mathrm{var}(S_n)^{1/2}) - \Phi(x/\mathrm{var}^*(S_n^\flat)^{1/2})\right| \leqslant \frac{1}{2}\left|\mathrm{var}^*(S_n^\flat)/\mathrm{var}(S_n) - 1\right| \tag{4.35}$$

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

$$\Phi(b/\mathrm{var}(S_n)^{1/2}) - \Phi(a/\mathrm{var}(S_n)^{1/2}) \leqslant \frac{b-a}{\sqrt{2\pi\,\mathrm{var}(S_n)}} \leqslant \frac{\bar{f}^{1/2}(b-a)}{\|\mathbf{S}^{-1/2}\boldsymbol{\lambda}\|_2\sqrt{2\pi\tau(1-\tau)}}. \tag{4.36}$$

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

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

where steps (i) and (iv) follow respectively from the Berry–Esseen inequalities (4.30) and (4.31), step (ii) uses the anti-concentration inequality (4.36) and step (iii) is due to the Gaussian comparison inequality (4.35). Conditioned on $\mathcal{E}_n \cap \mathcal{G}_n$,

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

Moreover, on the event $\mathcal{E}_n'$, the bounds in (4.34) 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 $\mathcal{E}_n \cap \mathcal{E}_n' \cap \mathcal{G}_n$,

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

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

Funding

National Science Foundation Award (DMS-1811376).

Acknowledgements

The authors are grateful to the associate editor and reviewers for thoughtful feedback and constructive comments. The second author would also like to thank Lan Wang and Jelena Bradic for helpful discussions and encouragement.

References

1. Adamczak, R. (2008) A tail inequality for suprema of unbounded empirical processes with applications to Markov chains. *Electron. J. Probab.*, **13**, 1000–1034.
2. Arcones, M. A. (1996) The Bahadur–Kiefer representation of L_p regression estimators. *Econ. Theory*, **12**, 257–283.
3. Arcones, M. A. & Giné, E. (1992) On the bootstrap of M-estimators and other statistical functionals. *Exploring the Limits of Bootstrap* (R. Le Page & L. Billard eds). New York: Wiley, pp. 14–47.
4. Bassett, G. & Koenker, R. (1978) Asymptotic theory of least absolute error regression. *J. Amer. Statist. Assoc.*, **73**, 618–622.
5. Bassett, G. & Koenker, R. (1986) Strong consistency of regression quantiles and related empirical processes. *Econ. Theory*, **2**, 191–201.
6. Boucheron, S., Bousquet, O., Lugosi, G. & Massart, P. (2005) Moment inequalities for functions of independent random variables. *Ann. Probab.*, **33**, 514–560.
7. Chatterjee, S. & Bose, A. (2005) Generalized bootstrap for estimating equations. *Ann. Stat.*, **33**, 414–436.
8. Chen, K., Ying, Z., Zhang, H. & Zhao, L. (2008) Analysis of least absolute deviation. *Biometrika*, **95**, 107–122.
9. Chen, X. & Zhou, W.-X. (2019) Robust inference via multiplier bootstrap. *Ann. Stat.*, to appear. Preprint arXiv:1903.07208.
10. Cheng, G. & Huang, J. Z. (2010) Bootstrap consistency for general semiparametric M-estimation. *Ann. Stat.*, **38**, 2884–2915.
11. Chernozhukov, V., Chetverikov, D. & Kato, K. (2014) Gaussian approximation of suprema of empirical processes. *Ann. Stat.*, **42**, 1564–1597.
12. Dudley, R. M. (1979) Balls in $\mathbf{r}^k$ do not cut all subsets of $k+2$ points. *Adv. Math.*, **31**, 306–308.
13. Efron, B. (1979) Bootstrap methods: another look at the jackknife. *Ann. Stat.*, **7**, 1–26.
14. Efron, B. & Tibshirani, R. J. (1994) *An Introduction to the Bootstrap*. New York: Chapman Hall.
15. Feng, X., He, X. & Hu, J. (2011) Wild bootstrap for quantile regression. *Biometrika*, **98**, 995–999.
16. Giné, E. & Zinn, J. (1990) Bootstrapping general empirical measures. *Ann. Probab.*, **18**, 851–869.
17. Gutenbrunner, C. & Jurečková, J. (1992) Regression rank scores and regression quantiles. *Ann. Stat.*, **20**, 305–330.

18. Gutenbrunner, C., Jurečková, J., Koenker, R. & Portnoy, S. (1993) Tests of linear hypotheses based on regression rank scores. *J. Nonparametr. Stat.*, **2**, 307–331.

19. He, X. & Hu, F. (2002) Markov chain marginal bootstrap. *J. Amer. Statist. Assoc.*, **97**, 783–795.

20. He, X. & Shao, Q.-M. (1996) A general Bahadur representation of M-estimators and its application to linear regression with nonstochastic designs. *Ann. Stat.*, **24**, 2608–2630.

21. Hsu, D., Kakade, S. M. & Zhang, T. (2014) Random design analysis of ridge regression. *Found. Comput. Math.*, **14**, 569–600.

22. Kocherginsky, M., He, X. & Mu, Y. (2005) Practical confidence intervals for regression quantiles. *J. Comp. Graph. Stat.*, **14**, 41–55.

23. Koenker, R. (1988) Asymptotic theory and econometric practice. *J. Appl. Econom.*, **3**, 139–147.

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

25. Koenker, R. (2019). Package 'quantreg', version 5.54. Manual: https://cran.r-project.org/web/packages/quantreg/quantreg.pdf.

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

27. Koenker, R. & Bassett, G. (1982) Tests of linear hypotheses and ℓ_1 estimation. *Econometrica*, **50**, 1577–1583.

28. Ledoux, M. & Talagrand, M. (1991) *Probability in Banach Spaces: Isoperimetry and Processes*. Ergebnisse der Mathematik und Ihrer Grenzgebiete (3), vol. 23. Berlin: Springer.

29. Loh, P.-L. & Wainwright, M. J. (2015) Regularized M-estimators with non-convexity: statistical and algorithmic theory for local optima. *J. Mach. Learn. Res.*, **16**, 559–616.

30. Ma, S. & Kosorok, M. R. (2005) Robust semiparametric M-estimation and the weighted bootstrap. *J. Multivariate Anal.*, **96**, 190–217.

31. McDiarmid, C. (1989) On the method of bounded differences. Surveys in Combinatorics, London Math. Soc. Lecture Note Ser., **141**. Cambridge: Cambridge University Press, pp. 148–188.

32. Negahban, S. N., Ravikumar, P., Wainwright, M. J. & Yu, B. (2012) A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Stat. Sci.*, **27**, 538–557.

33. Pan, X., Sun, Q. & Zhou, W.-X. (2019) Nonconvex regularized robust regression with oracle properties in polynomial time. Preprint arXiv:1907.04027.

34. Parzen, M. I., Wei, L. J. & Ying, Z. (1994) A resampling method based on pivotal estimating functions. *Biometrika*, **81**, 341–350.

35. Pollard, D. (1991) Asymptotics for least absolute deviation regression estimators. *Econ. Theory*, **7**, 186–199.

36. Portnoy, S. (1985) Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large; II. Normal approximation. *Ann. Stat.*, **13**, 1403–1417.

37. Portnoy, S. (1986) On the central limit theorem in R^p when $p \to \infty$. *Probab. Theory Relat. Fields*, **73**, 571–583.

38. Portnoy, S. & Koenker, R. (1989) Adaptive L-estimation for linear models. *Ann. Stat.*, **17**, 362–381.

39. Praestgaard, J. & Wellner, J. (1993) Exchangeably weighted bootstraps of the general empirical process. *Ann. Probab.*, **21**, 2053–2086.

40. Ruppert, D. & Carroll, R. J. (1980) Trimmed least squares estimation in the linear model. *J. Amer. Statist. Assoc.*, **75**, 828–838.

41. Silverman, B. (1986) *Density Estimation for Statistics and Data Analysis*. New York: Chapman Hall.

42. Spokoiny, V. & Zhilova, M. (2015) Bootstrap confidence sets under model misspecification. *Ann. Stat.*, **43**, 2653–2675.

43. Sun, Q., Zhou, W.-X. & Fan, J. (2019) Adaptive Huber regression. *J. Amer. Statist. Assoc.*, **115**, 254–265

44. Tyurin, I. S. (2011) On the convergence rate in Lyapunov's theorem. *Theory Probab. Appl.*, **55**, 253–270.

45. van de Geer, S. (2000) *Empirical Processes in M-Estimation*. Cambridge: Cambridge University Press.

46. van der Vaart, A. W. & Wellner, J. A. (1996) *Weak Convergence and Empirical Processes: With Applications to Statistics*. New York: Springer.

47. Wainwright, M. J. (2019) *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge: Cambridge University Press.
48. Wellner, J. A. & Zhan, Y. (1996) Bootstrapping *Z*-estimators. *Technical Report*. Seattle: Department of Statistics, University of Washington.
49. Welsh, A. H. (1989) On *M*-processes and *M*-estimation. *Ann. Stat.*, **15**, 337–361.
50. Zhao, L. C., Rao, C. R. & Chen, X. R. (1993) A note on the consistency of *M*-estimates in linear models. *Stochastic Processes: A Festschrift in Honour of Gopinath Kallianpur*. New York: Springer, pp. 359–367.

A. Proofs of Propositions 4.1– 4.4

A.1 *Proof of Proposition 4.1*

By (4.3), every $\boldsymbol{\xi}_{\boldsymbol{\beta}^*} \in \partial Q_n(\boldsymbol{\beta}^*)$ satisfies $\boldsymbol{\xi}_{\boldsymbol{\beta}^*} = \boldsymbol{\xi}^* := (1/n)\sum_{i=1}^n \boldsymbol{x}_i\{I(\varepsilon_i \leqslant 0) - \tau\}$ with probability one. Hence, it suffices to bound $\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\xi}^*\|_2 = \sup_{\|\boldsymbol{u}\|_2=1}\langle \boldsymbol{u}, \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\xi}^*\rangle$. Via a standard covering argument, for any $\epsilon \in (0,1)$, there exists an ϵ-net $\mathcal{N}_\epsilon$ of the unit sphere $\mathbb{S}^{d-1}$ with $|\mathcal{N}_\epsilon| \leqslant (1+2/\epsilon)^d$ such that $\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\xi}^*\|_2 \leqslant (1-\epsilon)^{-1}\max_{\boldsymbol{u}\in\mathcal{N}_\epsilon}\langle \boldsymbol{u}, \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\xi}^*\rangle$. Along each direction $\boldsymbol{u}$, define one-dimensional marginals

$$\gamma_{\boldsymbol{u},i} = \langle \boldsymbol{u}, \boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}_i\rangle\{I(\varepsilon_i \leqslant 0) - \tau\}, \quad i = 1, \ldots, n,$$

which satisfy $\mathbb{E}(\gamma_{\boldsymbol{u},i}) = 0$ and $\operatorname{var}(\gamma_{\boldsymbol{u},i}) = \tau(1-\tau) \leqslant 1/4$. By Condition 1, $\mathbb{P}(|\langle \boldsymbol{u}, \boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}_i\rangle| \geqslant \upsilon_0 t) \leqslant 2e^{-t^2/2}$ for all $t \geqslant 0$. Hence, for $k = 1, 2, \ldots$,

$$\begin{aligned}\mathbb{E}\gamma_{\boldsymbol{u},i}^{2k} &\leqslant \mathbb{E}\{I(\varepsilon_i \leqslant 0) - \tau\}^2\langle \boldsymbol{u}, \boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}_i\rangle^{2k} \\ &= \frac{1}{4}\upsilon_0^{2k}\, 2k\int_0^\infty \mathbb{P}\big(|\langle \boldsymbol{u}, \boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}_i\rangle| \geqslant \upsilon_0 t\big)t^{2k-1}\mathrm{d}t \leqslant \upsilon_0^{2k}\, 2^{k-1}k! \leqslant \frac{(2k)!}{2^k k!}(a_1\upsilon_0)^{2k}\end{aligned}$$

for some absolute constant $a_1 > 1$. Following the proof of Theorem 2.6 in Wainwright (2019), it can be shown that $\mathbb{E}e^{\lambda\gamma_{\boldsymbol{u},i}} \leqslant e^{(a_1a_2\lambda\upsilon_0)^2/2}$ for all $\lambda \in \mathbb{R}$, where $a_2 > 1$ is also an absolute constant. By the Hoeffding bound for sums of sub-Gaussian random variables (see, e.g. Proposition 2.5. in Wainwright, 2019), for any $y \geqslant 0$ we have

$$\frac{1}{n}\sum_{i=1}^n \gamma_{\boldsymbol{u},i} \leqslant a_1a_2\upsilon_0\sqrt{\frac{2y}{n}}$$

with probability at least $1 - e^{-y}$. Taking the union bound over all vectors $\boldsymbol{u} \in \mathcal{N}_\epsilon$ yields

$$\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\xi}^*\|_2 \leqslant \frac{a_1a_2\upsilon_0}{1-\epsilon}\sqrt{\frac{2y}{n}}$$

with probability greater than $1 - e^{\log(1+2/\epsilon)d - y}$. Through a careful analysis, we select $a_1 = 1.09$, $a_2 = 1.3$ and $\epsilon = 0.314$ so that all the requirements are satisfied. Finally, taking $y = 2d + x$ completes the proof.

A.2 *Proof of Proposition 4.2*

By (4.2), every $\boldsymbol{\xi}_{\boldsymbol{\beta}} = (\xi_{\boldsymbol{\beta},1}, \ldots, \xi_{\boldsymbol{\beta},d})^{\mathsf{T}} \in \partial Q_n(\boldsymbol{\beta})$ can be written as

$$\xi_{\boldsymbol{\beta},j} = -\frac{\tau}{n}\sum_{i=1}^{n} x_{ij} + \frac{1}{n}\sum_{i=1}^{n} x_{ij} I(y_i \leqslant \langle \boldsymbol{x}_i, \boldsymbol{\beta}\rangle) - \frac{1}{n}\sum_{i=1}^{n} x_{ij}\{v_i + (1-\tau)\} I(y_i = \langle \boldsymbol{x}_i, \boldsymbol{\beta}\rangle),$$

where $v_i \in [\tau - 1, \tau]$. With $\boldsymbol{\delta} = \boldsymbol{\beta} - \boldsymbol{\beta}^*$, it follows that

$$\begin{aligned}\langle \boldsymbol{\xi}_{\boldsymbol{\beta}} - \boldsymbol{\xi}_{\boldsymbol{\beta}^*}, \boldsymbol{\beta} - \boldsymbol{\beta}^*\rangle \geqslant \underbrace{\frac{1}{n}\sum_{i=1}^{n} \langle \boldsymbol{x}_i, \boldsymbol{\delta}\rangle \big\{ I(\varepsilon_i \leqslant \langle \boldsymbol{x}_i, \boldsymbol{\delta}\rangle) - I(\varepsilon_i \leqslant 0)\big\}}_{:= U_n(\boldsymbol{\delta})} \\ - \frac{1}{n}\sum_{i=1}^{n} |\langle \boldsymbol{x}_i, \boldsymbol{\delta}\rangle| \big\{ I(\varepsilon_i = \langle \boldsymbol{x}_i, \boldsymbol{\delta}\rangle) + I(\varepsilon_i = 0)\big\}. \end{aligned} \tag{A.1}$$

Since the conditional distribution of ε given $\boldsymbol{x}$ is continuous, with probability one, there is no vector $\boldsymbol{\delta} \in \mathbb{R}^d$ and $1 \leqslant i \leqslant n$ such that $\varepsilon_i = \langle \boldsymbol{x}_i, \boldsymbol{\delta}\rangle$. See Lemma A.1 of Ruppert & Carroll (1980). In other words, with probability one,

$$\frac{1}{n}\sum_{i=1}^{n} |\langle \boldsymbol{x}_i, \boldsymbol{\delta}\rangle| \big\{ I(\varepsilon_i = \langle \boldsymbol{x}_i, \boldsymbol{\delta}\rangle) + I(\varepsilon_i = 0)\big\} = 0 \text{ for all } \boldsymbol{\delta} \in \mathbb{R}^d. \tag{A.2}$$

Turning to the first term on the right-hand side of (A.1), the main difficulty comes from the discontinuity of $U_n(\boldsymbol{\delta})$ as a function of $\boldsymbol{\delta}$. To construct a smooth version of U_n, we introduce four Lipschitz continuous functions as follows. For any $a, b > 0$ and $u \in \mathbb{R}$, define

$$\varphi_a^+(u) = \begin{cases} 1 & \text{if } u > 2a \\ -1 + \frac{u}{a} & \text{if } a < u \leqslant 2a \\ 0 & \text{otherwise} \end{cases}, \quad \varphi_a^-(u) = \begin{cases} 1 & \text{if } u < -2a \\ -1 - \frac{u}{a} & \text{if } -2a \leqslant u < -a \\ 0 & \text{otherwise} \end{cases}, \tag{A.3}$$

and

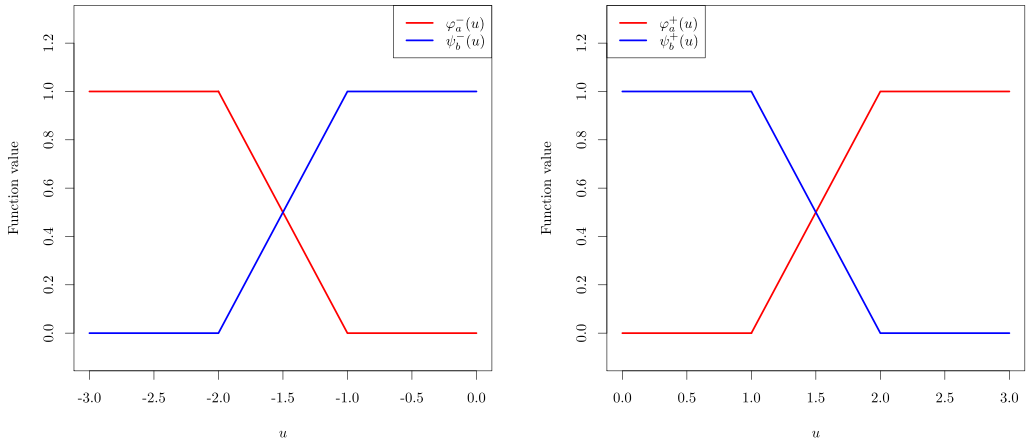
$$\psi_b^+(u) = \begin{cases} 1 & \text{if } u \leqslant b/2 \\ 2 - \frac{2u}{b} & \text{if } \frac{b}{2} < u \leqslant b \\ 0 & \text{otherwise} \end{cases}, \quad \psi_b^-(u) = \begin{cases} 1 & \text{if } u \geqslant -b/2 \\ 2 + \frac{2u}{b} & \text{if } -b \leqslant u < -\frac{b}{2} \\ 0 & \text{otherwise} \end{cases}. \tag{A.4}$$

Respectively, $\varphi_a^{\pm}$ and $\psi_b^{\pm}$ are $(1/a)$- and $(2/b)$-Lipschitz continuous, see Fig. A2. Also, they satisfy the following properties: for $a, b > 0$ and $u \in \mathbb{R}$,

$$I(u \geq 2a) \leqslant \varphi_a^+(u) \leqslant I(u \geq a), \quad I(u \leq -2a) \leqslant \varphi_a^-(u) \leqslant I(u \leq -a), \tag{A.5}$$

$$I(u \leq b/2) \leqslant \psi_b^+(u) \leqslant I(u \leqslant b), \quad I(u \geq -b/2) \leqslant \psi_b^-(u) \leqslant I(u \geq -b), \tag{A.6}$$

$$a\varphi_a^+(u) \leqslant \frac{1}{2}\max\{u, 0\}, \quad a\varphi_a^-(u) \leqslant \frac{1}{2}\max\{-u, 0\}. \tag{A.7}$$



(a) Plots of $\varphi_a^-(u)$ and $\psi_b^-(u)$. (b) Plots of $\varphi_a^+(u)$ and $\psi_b^+(u)$.

FIG. A2. The Lipschitz continuous functions $\varphi_a^{\pm}(u)$ and $\psi_b^{\pm}(u)$ with $a = 1$ and $b = 2$.

Furthermore, for each ε_i, we define its positive and negative components as $\varepsilon_{i,+} = \max\{\varepsilon_i, 0\}$ and $\varepsilon_{i,-} = \max\{-\varepsilon_i, 0\}$. For any $r > 0$, taking $a = \varepsilon_{i,\pm}$ and $b = 2r\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}}$ yields

$$
\begin{aligned}
U_n(\boldsymbol{\delta}) &= \frac{1}{n}\sum_{i=1}^{n}\left\{\langle \boldsymbol{x}_i, \boldsymbol{\delta}\rangle I(0 < \varepsilon_i \leqslant \langle \boldsymbol{x}_i, \boldsymbol{\delta}\rangle) + \langle -\boldsymbol{x}_i, \boldsymbol{\delta}\rangle I(\langle \boldsymbol{x}_i, \boldsymbol{\delta}\rangle < \varepsilon_i \leq 0)\right\} \\
&\geq \frac{1}{n}\sum_{i=1}^{n}\left\{\varepsilon_{i,+}\varphi^+_{\varepsilon_{i,+}}(\langle \boldsymbol{x}_i, \boldsymbol{\delta}\rangle) + \varepsilon_{i,-}\varphi^-_{\varepsilon_{i,-}}(\langle \boldsymbol{x}_i, \boldsymbol{\delta}\rangle)\right\} \\
&\geq \frac{1}{n}\sum_{i=1}^{n}\varepsilon_{i,+}\varphi^+_{\varepsilon_{i,+}}(\langle \boldsymbol{x}_i, \boldsymbol{\delta}\rangle) I(\langle \boldsymbol{x}_i, \boldsymbol{\delta}\rangle \leq 2r\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}}) + \frac{1}{n}\sum_{i=1}^{n}\varepsilon_{i,-}\varphi^-_{\varepsilon_{i,-}}(\langle \boldsymbol{x}_i, \boldsymbol{\delta}\rangle) I(\langle \boldsymbol{x}_i, \boldsymbol{\delta}\rangle \geq -2r\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}}) \\
&\geq \underbrace{\frac{1}{n}\sum_{i=1}^{n}\varepsilon_{i,+}\varphi^+_{\varepsilon_{i,+}}(\langle \boldsymbol{x}_i, \boldsymbol{\delta}\rangle)\psi^+_{2r\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}}}(\langle \boldsymbol{x}_i, \boldsymbol{\delta}\rangle)}_{V_n^+(\boldsymbol{\delta})} + \underbrace{\frac{1}{n}\sum_{i=1}^{n}\varepsilon_{i,-}\varphi^-_{\varepsilon_{i,-}}(\langle \boldsymbol{x}_i, \boldsymbol{\delta}\rangle)\psi^-_{2r\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}}}(\langle \boldsymbol{x}_i, \boldsymbol{\delta}\rangle)}_{V_n^-(\boldsymbol{\delta})}. \qquad \text{(A.8)}
\end{aligned}
$$

To bound $V_n(\boldsymbol{\delta}) = V_n^+(\boldsymbol{\delta}) + V_n^-(\boldsymbol{\delta})$ from below, we follow a two-step procedure: in step one, we derive a lower bound on the expectation $\mathbb{E}\{V_n(\boldsymbol{\delta})\}$, and in step two, we show concentration of $V_n(\boldsymbol{\delta})$ around $\mathbb{E}\{V_n(\boldsymbol{\delta})\}$ uniformly over $\boldsymbol{\delta} \in \mathbb{R}^d$ with high probability.

STEP 1. Along each direction $\boldsymbol{\delta} \in \mathbb{R}^d \setminus \{\mathbf{0}\}$, define the one-dimensional marginal $\eta_{\boldsymbol{\delta}} = \langle \boldsymbol{x}, \boldsymbol{\delta} \rangle / \|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}}$ that satisfies $\mathbb{E}(\eta_{\boldsymbol{\delta}}^2) = 1$. Using the lower bounds of $\varphi_a^{\pm}$ and $\psi_b^{\pm}$ given in (A.5) and (A.6), we obtain

$$\mathbb{E}\{V_n^+(\boldsymbol{\delta})\} \geq \frac{1}{n} \sum_{i=1}^n \mathbb{E}\big\{\varepsilon_{i,+} I(2\varepsilon_{i,+} \leq \langle \boldsymbol{x}_i, \boldsymbol{\delta} \rangle \leq r\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}})\big\},$$

$$\mathbb{E}\{V_n^-(\boldsymbol{\delta})\} \geq \frac{1}{n} \sum_{i=1}^n \mathbb{E}\big\{\varepsilon_{i,-} I(-r\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}} \leq \langle \boldsymbol{x}_i, \boldsymbol{\delta} \rangle \leq -2\varepsilon_{i,-})\big\}.$$

Together, with Condition 2 and the law of total expectation, we have

$$\begin{aligned}
\mathbb{E}\{V_n(\boldsymbol{\delta})\} &\geqslant \frac{1}{n} \sum_{i=1}^n \mathbb{E} \int_0^{\langle \boldsymbol{x}_i, \boldsymbol{\delta} \rangle / 2} t f_{\varepsilon_i | \boldsymbol{x}_i}(t)\, \mathrm{d}t \cdot I(0 \leq \langle \boldsymbol{x}_i, \boldsymbol{\delta} \rangle \leq r\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}}) \\
&\quad + \frac{1}{n} \sum_{i=1}^n \mathbb{E} \int_{\langle \boldsymbol{x}_i, \boldsymbol{\delta} \rangle / 2}^0 (-t) f_{\varepsilon_i | \boldsymbol{x}_i}(t)\, \mathrm{d}t \cdot I(-r\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}} \leq \langle \boldsymbol{x}_i, \boldsymbol{\delta} \rangle \leq 0) \\
&\geq \frac{1}{4} \|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}}^2 \cdot \mathbb{E}\big\{f_{\varepsilon | \boldsymbol{x}}(0) \eta_{\boldsymbol{\delta}}^2 I(|\eta_{\boldsymbol{\delta}}| \leq r)\big\} - \frac{L_0}{24} \|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}}^3 \cdot \mathbb{E}\big\{|\eta_{\boldsymbol{\delta}}|^3 I(|\eta_{\boldsymbol{\delta}}| \leq r)\big\} \\
&\geqslant \left(\frac{\underline{f}}{4} - \frac{L_0}{24} r\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}}\right) \|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}}^2 \mathbb{E}\big\{\eta_{\boldsymbol{\delta}}^2 I(|\eta_{\boldsymbol{\delta}}| \leq r)\big\}. \qquad \text{(A.9)}
\end{aligned}$$

Under Condition 1, $\mathbb{P}(|\eta_{\boldsymbol{\delta}}/\upsilon_0| \geqslant t) \leqslant 2e^{-t^2/2}$ for all $t \geqslant 0$ and $\boldsymbol{\delta} \in \mathbb{R}^d$. Therefore,

$$\begin{aligned}
\mathbb{E}\big\{\eta_{\boldsymbol{\delta}}^2 I(|\eta_{\boldsymbol{\delta}}| > r)\big\} &= \left(\int_0^{r^2} + \int_{r^2}^{\infty}\right) \mathbb{P}\big\{\eta_{\boldsymbol{\delta}}^2 I(|\eta_{\boldsymbol{\delta}}| > r) > t\big\}\, \mathrm{d}t \\
&= 2\upsilon_0^2 \int_{r/\upsilon_0}^{\infty} \mathbb{P}(|\eta_{\boldsymbol{\delta}}/\upsilon_0| \geqslant t) t\, \mathrm{d}t + r^2 \mathbb{P}(|\eta_{\boldsymbol{\delta}}/\upsilon_0| > r/\upsilon_0) \\
&\leqslant 2r^2 e^{-(r/\upsilon_0)^2/2} + 4\upsilon_0^2 \int_{(r/\upsilon_0)^2/2}^{\infty} e^{-s} \mathrm{d}s = \big(2r^2 + 4\upsilon_0^2\big) e^{-(r/\upsilon_0)^2/2}.
\end{aligned}$$

Taking $r = 4\upsilon_0^2$ with $\upsilon_0 \geqslant 1$, it follows that $\mathbb{E}\{\eta_{\boldsymbol{\delta}}^2 I(|\eta_{\boldsymbol{\delta}}| \leqslant r)\} \geqslant 1 - \sup_{\upsilon_0 \geqslant 1}(32\upsilon_0^4 + 4\upsilon_0^2) e^{-8\upsilon_0^2} \geqslant 1 - 36e^{-8}$. Substituting this into (A.9) yields

$$\mathbb{E}\{V_n(\boldsymbol{\delta})\} \geqslant (2/9 - 8e^{-8}) \underline{f}\, \|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}}^2 \qquad \text{(A.10)}$$

for all $\boldsymbol{\delta} \in \mathbb{R}^d$ satisfying $0 \leqslant \|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}} \leqslant \underline{f}/(6L_0\upsilon_0^2)$, where $\underline{f}$ and L_0 are defined in Condition 2.

STEP 2. We prove the concentration of $V_n(\boldsymbol{\delta})$ around $\mathbb{E}\{V_n(\boldsymbol{\delta})\}$ uniformly over $\boldsymbol{\delta}$ via the peeling technique, which is widely used in empirical process theory (van de Geer, 2000). For some $\delta > 0$ to be specified, define $\Theta(\delta) = \{\boldsymbol{\delta} \in \mathbb{R}^d : \|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}} \geqslant \delta\} = \cup_{\ell=1}^{\infty} \Theta_\ell(\delta)$ with

$$\Theta_\ell(\delta) = \big\{\boldsymbol{\delta} \in \mathbb{R}^d : 2^{(\ell-1)/2}\delta \leqslant \|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}} \leqslant 2^{\ell/2}\delta\big\}, \quad \ell = 1, 2 \ldots.$$

For any $R \geqslant \delta$, define

$$\Delta_n(R) = f(\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n; R) := \sup_{\delta \leqslant \|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}} \leqslant R} \big\{\mathbb{E}V_n(\boldsymbol{\delta}) - V_n(\boldsymbol{\delta})\big\}, \qquad \text{(A.11)}$$

where $\boldsymbol{w}_i = (\boldsymbol{x}_i, \varepsilon_i) \in \mathbb{R}^d \times \mathbb{R}$. For $\boldsymbol{\delta} \in \mathbb{R}^d$, write

$$\mathcal{E}(\boldsymbol{\delta}; \boldsymbol{w}_i) = \varepsilon_{i,+}\varphi^+_{\varepsilon_{i,+}}(\langle \boldsymbol{x}_i, \boldsymbol{\delta}\rangle)\psi^+_{2r\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}}}(\langle \boldsymbol{x}_i, \boldsymbol{\delta}\rangle) + \varepsilon_{i,-}\varphi^-_{\varepsilon_{i,-}}(\langle \boldsymbol{x}_i, \boldsymbol{\delta}\rangle)\psi^-_{2r\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}}}(\langle \boldsymbol{x}_i, \boldsymbol{\delta}\rangle).$$

Note that for any $b > 0$ and $u \in \mathbb{R}$, at most one of $\varepsilon_{i,+}\varphi^+_{\varepsilon_{i,+}}(u)\psi^+_b(u)$ and $\varepsilon_{i,-}\varphi^-_{\varepsilon_{i,-}}(u)\psi^-_b(u)$ can be non-zero. When $\langle \boldsymbol{x}_i, \boldsymbol{\delta}\rangle \geqslant 0$, by (A.4) and (A.7), we have

$$\begin{aligned} 0 \leqslant \mathcal{E}(\boldsymbol{\delta}; \boldsymbol{w}_i) &= \varepsilon_{i,+}\varphi^+_{\varepsilon_{i,+}}(\langle \boldsymbol{x}_i, \boldsymbol{\delta}\rangle)\psi^+_{2r\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}}}(\langle \boldsymbol{x}_i, \boldsymbol{\delta}\rangle) \\ &\leqslant \frac{\langle \boldsymbol{x}_i, \boldsymbol{\delta}\rangle}{2} \times \begin{cases} 1 & \text{if } \langle \boldsymbol{x}_i, \boldsymbol{\delta}\rangle \leqslant r\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}} \\ 2 - \frac{\langle \boldsymbol{x}_i, \boldsymbol{\delta}\rangle}{r\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}}} & \text{if } r\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}} < \langle \boldsymbol{x}_i, \boldsymbol{\delta}\rangle \leqslant 2r\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}} \\ 0 & \text{otherwise} \end{cases} \\ &\leqslant \frac{r}{2}\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}}. \end{aligned}$$

Following a similar argument, the same upper bound applies to $\mathcal{E}(\boldsymbol{\delta}; \boldsymbol{w}_i)$ when $\langle \boldsymbol{x}_i, \boldsymbol{\delta}\rangle < 0$. Consequently, we have $|\mathcal{E}(\boldsymbol{\delta}; \boldsymbol{w}_i)| \leqslant Rr/2$, so that for any index i and an independent copy $\boldsymbol{w}'_i = (\boldsymbol{x}'_i, \varepsilon'_i)$ of $\boldsymbol{w}_i = (\boldsymbol{x}_i, \varepsilon_i)$,

$$|f(\boldsymbol{w}_1, \ldots, \boldsymbol{w}_i, \ldots, \boldsymbol{w}_n; R) - f(\boldsymbol{w}_1, \ldots, \boldsymbol{w}'_i, \ldots, \boldsymbol{w}_n; R)| \leq \frac{Rr}{n}.$$

Hence, applying McDiarmid's inequality (McDiarmid, 1989), we obtain that for any $t \geqslant 0$,

$$\Delta_n(R) \leqslant \mathbb{E}\Delta_n(R) + Rr\sqrt{\frac{t}{2n}} \tag{A.12}$$

with probability at least $1 - e^{-t}$. Next we evaluate $\mathbb{E}\Delta_n(R)$. Again, using (A.7) it can be shown that for any $a, b > 0$, the functions $u \mapsto a\varphi^{\pm}_a(u)\psi^{\pm}_b(u)$ are 1-Lipschitz continuous. Thus, for any sample $\boldsymbol{w}_i = (\boldsymbol{x}_i, \varepsilon_i) \in \mathbb{R}^d \times \mathbb{R}$ and parameters $\boldsymbol{\delta}, \boldsymbol{\delta}' \in \mathbb{R}^d$, we have

$$|\mathcal{E}(\boldsymbol{\delta}; \boldsymbol{w}_i) - \mathcal{E}(\boldsymbol{\delta}'; \boldsymbol{w}_i)| \leqslant 2|\langle \boldsymbol{x}_i, \boldsymbol{\delta}\rangle - \langle \boldsymbol{x}_i, \boldsymbol{\delta}'\rangle|.$$

In other words, $\mathcal{E}(\boldsymbol{\delta}; \boldsymbol{w}_i)$ is a 2-Lipschitz continuous function in $\langle \boldsymbol{x}_i, \boldsymbol{\delta}\rangle$. Let $e_1, \ldots, e_n$ be independent Rademacher variables that are independent of the initial sample. By a classical symmetrization argument and the Ledoux–Talagrand contraction inequality (see, e.g. (4.20) in Ledoux & Talagrand, 1991),

$$\begin{aligned} \mathbb{E}\Delta_n(R) &\leqslant 2\mathbb{E}\left\{ \sup_{\delta \leqslant \|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}} \leqslant R} \frac{1}{n}\sum_{i=1}^n e_i \mathcal{E}(\boldsymbol{\delta}; \boldsymbol{w}_i) \right\} \leqslant 4\mathbb{E}\left\{ \sup_{\delta \leqslant \|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}} \leqslant R} \frac{1}{n}\sum_{i=1}^n e_i \langle \boldsymbol{x}_i, \boldsymbol{\delta}\rangle \right\} \\ &\leqslant 4R\,\mathbb{E}\left\| \frac{1}{n}\sum_{i=1}^n e_i \boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}_i \right\|_2 \leqslant 4R\sqrt{\frac{d}{n}}. \end{aligned} \tag{A.13}$$

Combining (A.12) and (A.13) yields that, with probability at least $1 - e^{-t}$,

$$\Delta_n(R) \leqslant Rr\sqrt{\frac{t}{2n}} + 4R\sqrt{\frac{d}{n}}. \tag{A.14}$$

With the above preparations, we derive that for any $t_0 > 0$,

$$\begin{aligned}
&\mathbb{P}\left\{\exists\, \boldsymbol{\delta} \in \Theta(\delta) \text{s.t.} - V_n(\boldsymbol{\delta}) + \mathbb{E}V_n(\boldsymbol{\delta}) \geqslant \|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}}^2 r t_0 + 4\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}}\sqrt{\frac{2d}{n}}\right\} \\
&\overset{\text{(i)}}{\leqslant} \sum_{\ell=1}^{\infty} \mathbb{P}\left\{\exists\, \boldsymbol{\delta} \in \Theta_\ell(\delta) \text{s.t.} - V_n(\boldsymbol{\delta}) + \mathbb{E}V_n(\boldsymbol{\delta}) \geqslant \frac{1}{2}(2^{\ell/2}\delta)^2 r t_0 + 4(2^{\ell/2}\delta)\sqrt{\frac{d}{n}}\right\} \\
&\overset{\text{(ii)}}{\leqslant} \sum_{\ell=1}^{\infty} \mathbb{P}\left\{\Delta_n(2^{\ell/2}\delta) \geqslant (2^{\ell/2}\delta) r\sqrt{\frac{(2^{\ell/2}t_0\delta)^2}{4}} + 4(2^{\ell/2}\delta)\sqrt{\frac{d}{n}}\right\} \\
&\overset{\text{(iii)}}{\leqslant} \sum_{\ell=1}^{\infty} e^{-(2^{\ell/2}t_0\delta)^2 n/2} = \sum_{\ell=1}^{\infty} e^{-2^{\ell-1}(t_0\delta)^2 n} \\
&\overset{\text{(iv)}}{\leqslant} \sum_{\ell=1}^{\infty} e^{-\ell(t_0\delta)^2 n} = \frac{e^{-(t_0\delta)^2 n}}{1 - e^{-(t_0\delta)^2 n}} := P(n, t_0, \delta),
\end{aligned} \tag{A.15}$$

where step (i) uses the union bound along with the decomposition $\Theta(\delta) = \cup_{\ell=1}^{\infty}\Theta_\ell(\delta)$, step (ii) follows from the definition of $\Delta_n(\cdot)$ in (A.11), step (iii) uses the concentration inequality (A.14) with $R = 2^{\ell/2}\delta$ for each $\ell \geqslant 1$ and step (iv) uses the elementary inequality that $2^{\ell-1} \geqslant \ell$.

Putting (A.1), (A.2), (A.8), (A.10) and (A.15) (with $r = 4\upsilon_0^2$) together, we conclude that with probability at least $1 - P(n, t_0, \delta)$,

$$\langle \boldsymbol{\xi}_{\boldsymbol{\beta}} - \boldsymbol{\xi}_{\boldsymbol{\beta}^*}, \boldsymbol{\beta} - \boldsymbol{\beta}^* \rangle \geqslant \left\{(2/9 - 8e^{-8})\underline{f} - 4\upsilon_0^2 t_0\right\}\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}}^2 - 4\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}}\sqrt{\frac{2d}{n}} \tag{A.16}$$

uniformly over $\boldsymbol{\delta} = \boldsymbol{\beta} - \boldsymbol{\beta}^*$ satisfying $\delta \leqslant \|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}} \leqslant \underline{f}/(6L_0\upsilon_0^2)$. In particular, we take $t_0 = (2/9 - 8e^{-8} - 1/8)\underline{f}/(4\upsilon_0^2)$ and recall that $\upsilon_0 \geq 1$ from Condition 1, then the right-hand side of (A.16) is bounded from below by

$$\frac{1}{8}\underline{f}\,\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}}^2 - 4\upsilon_0^2\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}}\sqrt{\frac{2d}{n}}.$$

By the convexity of Q_n, $\langle \boldsymbol{\xi}_{\boldsymbol{\beta}} - \boldsymbol{\xi}_{\boldsymbol{\beta}^*}, \boldsymbol{\beta} - \boldsymbol{\beta}^* \rangle$ is always non-negative. Therefore, for any $t \geqslant 0$, we may assume

$$\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma}} \geqslant \delta := (32\upsilon_0^2/\underline{f})\sqrt{\frac{2(d+t)}{n}};$$

otherwise, (4.4) holds trivially. The above choices of (t_0, δ) guarantee that $P(n, t_0, \delta) \leqslant e^{-t}/2$ in (A.15). Putting together the pieces, we conclude that with probability at least $1 - e^{-t}/2$,

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

for all $\boldsymbol{\beta}$ satisfying $0 \leqslant \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma}} \leqslant \underline{f}/(6L_0\upsilon_0^2)$. This completes the proof.

A.3 *Proof of Proposition 4.3*

By (4.3) and (4.5), every subgradient $\boldsymbol{\xi}^\flat = (\xi_1^\flat, \ldots, \xi_d^\flat)^\intercal \in \partial Q_n^\flat(\boldsymbol{\beta}^*)$ coincides with $(1/n)\sum_{i=1}^n w_i\zeta_i\boldsymbol{x}_i$ with probability one. Thus, without loss of generality, we assume $\boldsymbol{\xi}^\flat = (1/n)\sum_{i=1}^n w_i\zeta_i\boldsymbol{x}_i$. Note that

$\mathbb{E}^*\boldsymbol{\xi}^\flat = (1/n)\sum_{i=1}^n \zeta_i \boldsymbol{x}_i$. Using a standard covering argument again, for any $\epsilon \in (0,1)$, there exists an ϵ-net $\mathcal{N}_\epsilon \subseteq \mathbb{S}^{d-1}$ with $|\mathcal{N}_\epsilon| \leqslant (1+2/\epsilon)^d$ such that

$$\|\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\xi}^\flat - \mathbb{E}^*\boldsymbol{\xi}^\flat)\|_2 \leq \frac{1}{1-\epsilon}\max_{\boldsymbol{u}\in\mathcal{N}_\epsilon}\frac{1}{n}\sum_{i=1}^n e_i\zeta_i\langle \boldsymbol{u}, \boldsymbol{z}_i\rangle,$$

where e_i are independent Rademacher random variables. For any $\boldsymbol{u} \in \mathcal{N}_\epsilon$ and $y \geqslant 0$, by Hoeffding's inequality we have

$$\mathbb{P}^*\left\{\frac{1}{n}\sum_{i=1}^n e_i\zeta_i\langle \boldsymbol{u}, \boldsymbol{z}_i\rangle \geqslant \left(2y\sum_{i=1}^n \zeta_i^2\langle \boldsymbol{u}, \boldsymbol{z}_i\rangle^2\right)^{1/2}\frac{1}{n}\right\} \leqslant e^{-y}.$$

Moreover, note that ζ_i are bounded random variables that satisfy $\mathbb{E}(\zeta_i^2|\boldsymbol{x}_i) = \tau(1-\tau) \leqslant 1/4$, $\mathbb{E}(\zeta_i^4|\boldsymbol{x}_i) \leqslant 1/12$ and $|\zeta_i| \leqslant 1$. Following the calculations as in the proof of Proposition 4.1, for every $\boldsymbol{u} \in \mathcal{N}_\epsilon$ we have

$$\mathbb{E}(\zeta_i^2\langle \boldsymbol{u}, \boldsymbol{z}_i\rangle^2)^k \leqslant \frac{k!}{2}\frac{4}{3}\upsilon_0^4(2\upsilon_0^2)^{k-2}, \quad k = 2, 3, \ldots.$$

Using Bernstein's inequality, we obtain

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n \zeta_i^2\langle \boldsymbol{u}, \boldsymbol{z}_i\rangle^2 \geqslant \frac{1}{4} + 2\upsilon_0^2\sqrt{\frac{2x}{3n}} + 2\upsilon_0^2\frac{x}{n}\right) \leqslant e^{-x}, \text{ valid for any } x \geqslant 0.$$

Finally, we set $\epsilon = 2/(e^2-1)$ so that $(1+2/\epsilon)^d = e^{2d}$. Taking the union bound twice over all $\boldsymbol{u} \in \mathcal{N}_\epsilon$ with $x = y = 2(d+t)$ yields

$$\max_{\boldsymbol{u}\in\mathcal{N}_\epsilon}\frac{2}{n}\sum_{i=1}^n \zeta_i^2\langle \boldsymbol{u}, \boldsymbol{z}_i\rangle^2 \leqslant \frac{1}{2} + 8\upsilon_0^2\sqrt{\frac{d+t}{3n}} + 8\upsilon_0^2\frac{d+t}{n} \tag{A.17}$$

with probability at least $1-e^{-2t}$, and with $\mathbb{P}^*$-probability at least $1-e^{-2t}$ conditioned on the event that (A.17) holds,

$$\|\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\xi}^\flat - \mathbb{E}^*\boldsymbol{\xi}^\flat)\|_2 \leqslant 2\sqrt{\frac{d+t}{n}}$$

provided that $n \geqslant C\upsilon_0^4(d+t)$ for some universal constant $C > 0$. Putting together the pieces completes the proof of (4.6).

A.4 *Proof of Proposition 4.4*

We keep the notation used in the proof of Proposition 4.2 and follow a similar argument. To begin with, note that every $\boldsymbol{\xi}_{\boldsymbol{\beta}}^\flat = (\xi_{\boldsymbol{\beta},1}^\flat, \ldots, \xi_{\boldsymbol{\beta},d}^\flat)^\intercal \in \partial Q_n^\flat(\boldsymbol{\beta})$ can be written as

$$\xi_{\boldsymbol{\beta},j}^\flat = \frac{1}{n}\sum_{i=1}^n w_i x_{ij}\{I(y_i \leqslant \langle \boldsymbol{x}_i, \boldsymbol{\beta}\rangle) - \tau\} - \frac{1}{n}\sum_{i=1}^n w_i x_{ij}\{v_i + (1-\tau)\}I(y_i = \langle \boldsymbol{x}_i, \boldsymbol{\beta}\rangle),$$

where $v_i \in [\tau - 1, \tau]$. As before, the bound $\langle \boldsymbol{\xi}_{\boldsymbol{\beta}}^{\flat} - \boldsymbol{\xi}_{\boldsymbol{\beta}^*}^{\flat}, \boldsymbol{\beta} - \boldsymbol{\beta}^* \rangle \geqslant U_n^{\flat}(\boldsymbol{\delta})$ holds with probability one, where

$$U_n^{\flat}(\boldsymbol{\delta}) := \frac{1}{n} \sum_{i=1}^{n} w_i \langle \boldsymbol{x}_i, \boldsymbol{\delta} \rangle \{ I(\varepsilon_i \leqslant \langle \boldsymbol{x}_i, \boldsymbol{\delta} \rangle) - I(\varepsilon_i \leqslant 0) \} \text{ for } \boldsymbol{\delta} = \boldsymbol{\beta} - \boldsymbol{\beta}^*.$$

Again, introducing Lipschitz continuous functions $\varphi_a^{\pm}(u)$ and $\psi_b^{\pm}(u)$ as in (A.3) and (A.4), we obtain

$$\begin{aligned} U_n^{\flat}(\boldsymbol{\delta}) &\geqslant \frac{1}{n} \sum_{i=1}^{n} w_i \varepsilon_{i,+} \varphi_{\varepsilon_{i,+}}^{+}(\langle \boldsymbol{x}_i, \boldsymbol{\delta} \rangle) \psi_{2r\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}}}^{+}(\langle \boldsymbol{x}_i, \boldsymbol{\delta} \rangle) \\ &\quad + \frac{1}{n} \sum_{i=1}^{n} w_i \varepsilon_{i,-} \varphi_{\varepsilon_{i,-}}^{-}(\langle \boldsymbol{x}_i, \boldsymbol{\delta} \rangle) \psi_{2r\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}}}^{-}(\langle \boldsymbol{x}_i, \boldsymbol{\delta} \rangle) := V_n(\boldsymbol{\delta}) + V_n^{\flat}(\boldsymbol{\delta}), \end{aligned} \tag{A.18}$$

where $V_n(\boldsymbol{\delta}) = V_n^{+}(\boldsymbol{\delta}) + V_n^{-}(\boldsymbol{\delta})$ is defined in (A.8) and

$$V_n^{\flat}(\boldsymbol{\delta}) = \frac{1}{n} \sum_{i=1}^{n} e_i \varepsilon_{i,+} \varphi_{\varepsilon_{i,+}}^{+}(\langle \boldsymbol{x}_i, \boldsymbol{\delta} \rangle) \psi_{2r\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}}}^{+}(\langle \boldsymbol{x}_i, \boldsymbol{\delta} \rangle) + \frac{1}{n} \sum_{i=1}^{n} e_i \varepsilon_{i,-} \varphi_{\varepsilon_{i,-}}^{-}(\langle \boldsymbol{x}_i, \boldsymbol{\delta} \rangle) \psi_{2r\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}}}^{-}(\langle \boldsymbol{x}_i, \boldsymbol{\delta} \rangle).$$

Notice that $\mathbb{E}^*\{V_n^{\flat}(\boldsymbol{\delta})\} = 0$. For any $R \geqslant \delta$, define $\Gamma_n(R) = f(e_1, \ldots, e_n; R) := \sup_{\delta \leqslant \|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}} \leqslant R} - V_n^{\flat}(\boldsymbol{\delta})$. For each index i and an independent copy $\widetilde{e}_i$ of e_i, we have

$$|f(e_1, \ldots, e_i, \ldots, e_n; R) - f(e_1, \ldots, \widetilde{e}_i, \ldots, e_n; R)| \leq \frac{Rr}{n}.$$

Applying McDiarmid's inequality gives

$$\Gamma_n(R) \leq \mathbb{E}^*\{\Gamma_n(R)\} + Rr\sqrt{\frac{t}{2n}} \tag{A.19}$$

with $\mathbb{P}^*$-probability at least $1 - e^{-t}$. Using the Lipschitz continuity of $u \mapsto \varepsilon_{i,\pm} \varphi_{\varepsilon_{i,\pm}}^{+}(u) \psi_b^{\pm}(u)$, and Talagrand's contraction principle, we obtain

$$\begin{aligned} \mathbb{E}^*\{\Gamma_n(R)\} &\leqslant 2\mathbb{E}^*\left(\sup_{\delta \leqslant \|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}} \leqslant R} \frac{1}{n} \sum_{i=1}^{n} e_i \langle \boldsymbol{x}_i, \boldsymbol{\delta} \rangle \right) \leq \frac{2R}{n} \mathbb{E}^* \left\| \sum_{i=1}^{n} e_i \boldsymbol{z}_i \right\|_2 \\ &\leqslant \frac{2R}{n} \left(\sum_{i=1}^{n} \|\boldsymbol{z}_i\|_2^2 \right)^{1/2} = 2RM_{n,d}\sqrt{\frac{d}{n}}, \end{aligned} \tag{A.20}$$

where $\boldsymbol{z}_i$ are defined in (4.5) and $M_{n,d}^2 := (1/nd) \sum_{i=1}^{n} \sum_{j=1}^{d} z_{ij}^2$. Together, (A.19) and (A.20) imply

$$\Gamma_n(R) \leqslant 2RM_{n,d}\sqrt{\frac{d}{n}} + Rr\sqrt{\frac{t}{2n}} \tag{A.21}$$

with $\mathbb{P}^*$-probability at least $1 - e^{-t}$.

Note that inequality (A.21) holds for every $R \geqslant \delta$. Again, via the slicing technique and taking $r = 4\upsilon_0^2$, it can be shown that for any $t_1 > 0$, with $\mathbb{P}^*$-probability at least $1 - \frac{e^{-(t_1\delta)^2 n}}{1-e^{-(t_1\delta)^2 n}} = 1 - P(n, t_1, \delta)$,

$$V_n^\flat(\boldsymbol{\delta}) \geqslant -2M_{n,d}\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}}\sqrt{\frac{2d}{n}} - 4t_1\upsilon_0^2\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}}^2$$

uniformly over $\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}} \geq \delta$. For the data-dependent quantity $M_{n,d}$, note that $M_{n,d}^2 \leqslant \max_{1\leqslant j\leqslant d}(1/n)\sum_{i=1}^n z_{ij}^2$. Under Condition 1, we have $\mathbb{E}(z_{ij}^2) = 1$ and for $k = 2, 3, \ldots,$

$$\mathbb{E}(z_{ij}^2)^k = \upsilon_0^{2k}2k\int_0^\infty \mathbb{P}(|z_{ij}| \geqslant \upsilon_0 x)x^{2k-1}\mathrm{d}x \leqslant 2^{k+1}\upsilon_0^{2k}k! = \frac{k!}{2}16\upsilon_0^4(2\upsilon_0^2)^{k-2}.$$

It then follows from Bernstein's inequality that, for any $1 \leqslant j \leqslant d$ and $x \geqslant 0$,

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n z_{ij}^2 \geqslant 1 + 4\upsilon_0^2\sqrt{\frac{2x}{n}} + 2\upsilon_0^2\frac{x}{n}\right) \leqslant e^{-x}.$$

Taking $x = \log(2d) + t$ and applying the union bound, we obtain

$$M_{n,d}^2 \leq 1 + 4\upsilon_0^2\sqrt{\frac{2\log(2d) + 2t}{n}} + 2\upsilon_0^2\frac{\log(2d) + t}{n} \tag{A.22}$$

with probability at least $1 - e^{-t}/2$.

Turning to $V_n(\boldsymbol{\delta})$ in (A.18), it follows from (A.15) and (A.16) that with probability at least $1 - P(n, t_0, \delta)$,

$$V_n(\boldsymbol{\delta}) \geqslant \left\{(2/9 - 8e^{-8})\underline{f} - 4\upsilon_0^2 t_0\right\}\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}}^2 - 4\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}}\sqrt{\frac{2d}{n}} \tag{A.23}$$

for all $\boldsymbol{\delta}$ satisfying $\delta \leqslant \|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}} \leqslant \underline{f}/(6L_0\upsilon_0^2)$. Let $\mathcal{G}(t, t_0, \delta)$ be the event that (A.22) and (A.23) hold. Then $\mathbb{P}\{\mathcal{G}(t, t_0, \delta)\} \geqslant 1 - e^{-t}/2 - P(n, t_0, \delta)$. Taking $t_0 = t_1 = (2/9 - 8e^{-8} - 1/8)\underline{f}/(8\upsilon_0^2)$ yields that with $\mathbb{P}^*$-probability at least $1 - P(n, t_1, \delta)$ conditioned on $\mathcal{G}(t, t_0, \delta)$,

$$\langle \boldsymbol{\xi}_{\boldsymbol{\beta}}^\flat - \boldsymbol{\xi}_{\boldsymbol{\beta}^*}^\flat, \boldsymbol{\beta} - \boldsymbol{\beta}^*\rangle \geqslant \frac{1}{8}\underline{f}\,\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}}^2 - 8\upsilon_0^2\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}}\sqrt{\frac{2d}{n}}$$

uniformly over $\delta \leqslant \|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}} \leqslant \underline{f}/(6L_0\upsilon_0^2)$ as long as $n \geqslant C\upsilon_0^4\{\log(d) + t\}$ for some universal constant $C > 0$. For any $t \geqslant 0$, we assume that

$$\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma}} \geqslant \delta := (64\upsilon_0^2/\underline{f})\sqrt{\frac{2(d+t)}{n}};$$

otherwise, (4.7) holds trivially, and the above choices of (t_0, t_1, δ) guarantee that $P(n, t_0, \delta) = P(n, t_1, \delta) \leqslant e^{-t}/2$. This completes the proof.

B. Additional simulation studies

This section presents additional numerical results under various combinations of the design and error distributions.

B.1 *Confidence estimation*

B.2 *Goodness-of-fit testing*

TABLE B7 *Average coverage probabilities and CI widths over all the coefficients under homoscedastic model (3.1) with t_2 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	0.963	0.952	0.494	0.507	0.435	0.437	0.434
0.1:	0.935	0.945	0.905	0.923	0.909	0.414	0.425	0.365	0.361	0.364
0.2:	0.871	0.879	0.822	0.823	0.819	0.323	0.331	0.285	0.277	0.284
	Weakly correlated Gaussian design									
	Coverage probability					Width				
α	pair	pwy	wild	mb-per	mb-norm	pair	pwy	wild	mb-per	mb-norm
0.05:	0.978	0.980	0.951	0.970	0.955	0.735	0.754	0.650	0.652	0.646
0.1:	0.946	0.952	0.908	0.929	0.911	0.617	0.633	0.545	0.538	0.542
0.2:	0.871	0.876	0.812	0.836	0.820	0.481	0.493	0.425	0.412	0.422
	Equally correlated Gaussian design									
	Coverage probability					Width				
α	pair	pwy	wild	mb-per	mb-norm	pair	pwy	wild	mb-per	mb-norm
0.05:	0.980	0.982	0.950	0.974	0.958	0.784	0.804	0.693	0.694	0.688
0.1:	0.947	0.954	0.914	0.934	0.912	0.658	0.675	0.581	0.573	0.577
0.2:	0.871	0.881	0.821	0.837	0.820	0.512	0.526	0.453	0.438	0.450

TABLE B8 *Average coverage probabilities and CI widths over all the coefficients under heteroscedastic model (3.2) with t_2 error*

Independent Gaussian design										
	Coverage probability					Width				
α	pair	pwy	wild	mb-per	mb-norm	pair	pwy	wild	mb-per	mb-norm
0.05:	0.981	0.980	0.958	0.965	0.966	0.435	0.447	0.379	0.389	0.384
0.1:	0.958	0.959	0.919	0.920	0.926	0.365	0.375	0.318	0.320	0.323
0.2:	0.879	0.891	0.814	0.824	0.826	0.285	0.292	0.248	0.243	0.251
Weakly correlated Gaussian design										
	Coverage probability					Width				
α	pair	pwy	wild	mb-per	mb-norm	pair	pwy	wild	mb-per	mb-norm
0.05:	0.980	0.984	0.960	0.970	0.957	0.667	0.685	0.586	0.593	0.586
0.1:	0.952	0.957	0.922	0.926	0.920	0.560	0.575	0.492	0.488	0.492
0.2:	0.877	0.889	0.843	0.835	0.834	0.436	0.448	0.383	0.372	0.383
Equally correlated Gaussian design										
	Coverage probability					Width				
α	pair	pwy	wild	mb-per	mb-norm	pair	pwy	wild	mb-per	mb-norm
0.05:	0.986	0.990	0.968	0.974	0.974	0.716	0.734	0.626	0.636	0.629
0.1:	0.964	0.970	0.927	0.933	0.931	0.601	0.616	0.526	0.523	0.528
0.2:	0.887	0.898	0.835	0.836	0.840	0.468	0.480	0.409	0.399	0.411

TABLE B9 *Average coverage probabilities and CI widths over all the coefficients under homoscedastic model (3.1) with type II 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:	0.966	0.968	0.939	0.966	0.943	0.438	0.448	0.388	0.386	0.383
0.1:	0.929	0.936	0.889	0.920	0.891	0.367	0.376	0.326	0.320	0.322
0.2:	0.847	0.857	0.784	0.821	0.787	0.286	0.293	0.254	0.245	0.251
Weakly correlated Gaussian design										
	Coverage probability					Width				
α	pair	pwy	wild	mb-per	mb-norm	pair	pwy	wild	mb-per	mb-norm
0.05:	0.970	0.972	0.938	0.964	0.945	0.652	0.669	0.580	0.576	0.572
0.1:	0.935	0.939	0.886	0.920	0.891	0.547	0.562	0.487	0.477	0.480
0.2:	0.852	0.860	0.800	0.834	0.796	0.427	0.438	0.379	0.366	0.374
Equally correlated Gaussian design										
	Coverage probability					Width				
α	pair	pwy	wild	mb-per	mb-norm	pair	pwy	wild	mb-per	mb-norm
0.05:	0.962	0.967	0.937	0.966	0.941	0.695	0.713	0.618	0.614	0.610
0.1:	0.932	0.939	0.891	0.924	0.891	0.583	0.599	0.519	0.509	0.512
0.2:	0.850	0.859	0.803	0.825	0.801	0.454	0.466	0.404	0.390	0.399

TABLE B10 *Average coverage probabilities and CI widths over all the coefficients under heteroscedastic model (3.2) with type II 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:	0.977	0.980	0.956	0.966	0.961	0.386	0.397	0.339	0.344	0.340
0.1:	0.949	0.956	0.910	0.924	0.914	0.324	0.333	0.284	0.283	0.286
0.2:	0.870	0.879	0.816	0.821	0.818	0.253	0.259	0.222	0.216	0.223
	Weakly correlated Gaussian design									
	Coverage probability					Width				
α	pair	pwy	wild	mb-per	mb-norm	pair	pwy	wild	mb-per	mb-norm
0.05:	0.976	0.980	0.951	0.963	0.950	0.591	0.607	0.522	0.525	0.519
0.1:	0.941	0.946	0.901	0.922	0.908	0.496	0.509	0.438	0.433	0.436
0.2:	0.869	0.883	0.815	0.825	0.819	0.387	0.397	0.341	0.331	0.340
	Equally correlated Gaussian design									
	Coverage probability					Width				
α	pair	pwy	wild	mb-per	mb-norm	pair	pwy	wild	mb-per	mb-norm
0.05:	0.972	0.976	0.950	0.965	0.946	0.626	0.643	0.553	0.556	0.551
0.1:	0.937	0.946	0.903	0.926	0.908	0.525	0.540	0.464	0.459	0.462
0.2:	0.866	0.874	0.815	0.839	0.812	0.409	0.421	0.361	0.350	0.360

TABLE B11 *Average type I error and power under homoscedastic model (3.1) with t_2 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	0.330	0.005	0.000	0.005	0.945	0.490	0.535	0.600	0.685	0.095	0.075	0.085
0.05	0.465	0.015	0.035	0.020	0.985	0.735	0.865	0.855	0.790	0.330	0.330	0.320
0.1	0.550	0.070	0.080	0.055	0.990	0.845	0.945	0.930	0.840	0.445	0.505	0.440
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	0.320	0.000	0.000	0.005	0.785	0.295	0.365	0.370	0.835	0.350	0.375	0.405
0.05	0.475	0.040	0.020	0.015	0.900	0.595	0.645	0.595	0.940	0.650	0.710	0.665
0.1	0.565	0.065	0.050	0.040	0.935	0.715	0.750	0.730	0.975	0.770	0.815	0.805
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	0.320	0.000	0.000	0.005	0.810	0.330	0.345	0.380	0.985	0.760	0.850	0.845
0.05	0.475	0.040	0.020	0.015	0.890	0.570	0.630	0.595	0.995	0.935	0.965	0.965
0.1	0.565	0.065	0.050	0.040	0.935	0.690	0.755	0.735	1.000	0.980	0.990	0.990

TABLE B12 *Average type I error and power under heteroscedastic model (3.2) with t_2 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	0.325	0.005	0.000	0.000	0.935	0.650	0.770	0.810	0.745	0.205	0.160	0.200
0.05	0.475	0.020	0.010	0.005	0.975	0.865	0.935	0.940	0.840	0.480	0.425	0.395
0.1	0.520	0.055	0.045	0.020	0.990	0.940	0.965	0.960	0.890	0.620	0.630	0.555
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	0.315	0.000	0.000	0.000	0.840	0.415	0.490	0.510	0.905	0.470	0.475	0.535
0.05	0.465	0.050	0.010	0.010	0.910	0.685	0.705	0.685	0.955	0.750	0.785	0.775
0.1	0.550	0.095	0.030	0.025	0.945	0.810	0.815	0.790	0.975	0.845	0.900	0.865
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	0.015	0.000	0.000	0.845	0.425	0.450	0.500	0.990	0.840	0.920	0.930
0.05	0.445	0.040	0.015	0.005	0.925	0.675	0.750	0.745	0.995	0.970	0.985	0.980
0.1	0.510	0.085	0.060	0.030	0.950	0.800	0.870	0.835	0.995	0.990	1.000	1.000

TABLE B13 *Average type I error and power under homoscedastic model (3.1) with type II 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	0.290	0.005	0.015	0.015	0.945	0.610	0.680	0.705	0.695	0.150	0.175	0.180
0.05	0.435	0.040	0.035	0.030	0.990	0.800	0.895	0.900	0.860	0.395	0.495	0.440
0.1	0.505	0.095	0.090	0.055	1.000	0.910	0.935	0.920	0.890	0.595	0.645	0.595
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	0.345	0.005	0.005	0.005	0.860	0.420	0.530	0.565	0.905	0.455	0.530	0.525
0.05	0.525	0.025	0.035	0.030	0.945	0.695	0.760	0.725	0.945	0.675	0.775	0.725
0.1	0.600	0.075	0.110	0.085	0.965	0.815	0.875	0.855	0.960	0.810	0.860	0.830
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	0.345	0.005	0.005	0.005	0.875	0.385	0.485	0.525	0.980	0.890	0.950	0.940
0.05	0.525	0.025	0.035	0.030	0.930	0.670	0.725	0.700	1.000	0.980	0.995	0.990
0.1	0.600	0.075	0.110	0.085	0.960	0.810	0.845	0.790	1.000	0.990	0.995	0.995

TABLE B14 *Average type I error and power under heteroscedastic model (3.2) with type II 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	0.320	0.010	0.005	0.005	0.960	0.760	0.840	0.840	0.815	0.300	0.250	0.260
0.05	0.440	0.035	0.020	0.025	0.990	0.915	0.965	0.955	0.920	0.595	0.630	0.565
0.1	0.495	0.085	0.060	0.045	0.995	0.970	0.975	0.970	0.950	0.755	0.755	0.725
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	0.350	0.010	0.000	0.000	0.900	0.570	0.640	0.665	0.925	0.545	0.625	0.655
0.05	0.485	0.040	0.025	0.020	0.950	0.790	0.850	0.835	0.975	0.790	0.870	0.855
0.1	0.590	0.095	0.060	0.045	0.965	0.900	0.930	0.890	1.000	0.870	0.925	0.910
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	0.005	0.000	0.000	0.915	0.525	0.635	0.655	1.000	0.920	0.970	0.970
0.05	0.415	0.040	0.015	0.015	0.955	0.805	0.855	0.835	1.000	0.980	0.995	0.995
0.1	0.525	0.085	0.055	0.040	0.980	0.885	0.920	0.910	1.000	0.995	1.000	1.000