Sparse Composite Quantile Regression in Ultrahigh Dimensions with Tuning Parameter Calibration

Yuwen Gu, Hui Zou

Abstract

When estimating coefficients in a linear model, the (sparse) composite quantile regression was first proposed in Zou and Yuan (2008) as an efficient alternative to the (sparse) least squares regardless of the error distribution. The highly nonsmooth nature of the composite loss in the sparse composite quantile regression makes its theoretical analysis as well as numerical computation much more challenging than the least squares method. The theory in Zou and Yuan (2008) was proven under fixed-dimension asymptotics and the estimator was computed via linear programming. None of these can be extended to ultrahigh dimensions. In this paper, we study the sparse composite quantile regression under ultrahigh dimensionality and make three contributions. First, we provide a non-asymptotic analysis of both the lasso and the folded concave penalized composite quantile regression, which reveals a practical way of achieving the oracle estimator. Second, we construct a novel information criterion for selecting the regularization parameter in the folded concave penalized composite quantile regression and prove its selection consistency. Third, we exploit the structure of the composite loss and design a specialized optimization algorithm for computing the penalized composite quantile regression via the alternating direction method of multipliers. We conduct extensive simulations to illustrate the theoretical results. Our analysis provides a unified treatment of the concentration inequalities involving the composite loss. Those inequalities could be of independent interest.

I. Introduction

Coefficient estimation in linear models is routinely done via the least squares (LS) regression. Under Gaussian errors, the LS estimator has the likelihood interpretation and is most efficient. It is reasonably efficient under other light-tailed error distributions besides Gaussian. When the error distribution is heavy-tailed, the LS estimator may fail to be consistent. See numerical studies in Section VI for a clear demonstration. The quantile regression (QR, [1]) can consistently estimate the coefficients of a linear model under very heavy-tailed errors, like a Student's t (with few degrees of freedom) or Cauchy. The robustness of the QR estimator, a property often mentioned in the literature, comes from the fact that its asymptotic variance does not depend on the moments of the error distribution, upon which that of the LS estimator relies, however. In terms of efficiency, it is well known that the asymptotic

Y. Gu is with the Department of Statistics, University of Connecticut, Storrs, CT, USA (e-mail: yuwen.gu@uconn.edu).

H. Zou is with the School of Statistics, University of Minnesota, Minneapolis, MN, USA (e-mail: zouxx019@umn.edu).

Manuscript received XXXXXXXXXX; revised XXXXXXXXXX.

variance of a QR estimator is inversely proportional to (the square of) the error density evaluated at the true quantile of the error distribution ([2], [3]). Thus, under certain errors, it is expected that the QR estimator is more efficient than the LS estimator. Nevertheless, the quantile regression considers only one quantile at a time and may not fully grasp the distributional information to always produce efficient estimation. To its extreme, when the error density at a specified quantile approaches zero, the asymptotic variance of the corresponding QR estimator explodes to infinity, which results in an estimator having arbitrarily small efficiency. As an example, under the mixture normal error $\frac{1}{2}N(-3,1)+\frac{1}{2}N(3,1)$, the least absolute deviation estimator is 1272.8 times less efficient than the LS estimator.

To safeguard quantile regression against potential efficiency loss, methods based on the idea of combining quantile information across multiple levels have been proposed in the literature. The idea is natural: as more quantiles are used, we have more distributional information to dispense and can hence obtain more efficient estimation if we do it properly. One such approach named the composite quantile regression (CQR, [4], [5]) combines information over different quantiles via a mix of quantile loss functions. It was shown by [4] that the CQR estimator is much more (or arbitrarily more) efficient than the LS estimator under many heavy-tailed errors. Another notable approach by [6] seeks an optimal weighting scheme to combine QR estimators at given levels to achieve as much efficiency gain as possible. It was shown that as the number of quantiles increases, the asymptotic variance of their proposed estimator achieves the Cramér–Rao lower bound under certain regularity conditions.

When considering fitting a sparse CQR model, it is natural to adopt the sparse penalties used in the sparse LS. In [4], Zou and Yuan studied the sparse CQR using the adaptive lasso penalty [7] and proved its oracle properties under fixed-dimension asymptotics. We note that the approach by [6] cannot be easily regularized to obtain desired sparse solutions since their estimator is based on a weighted average of multiple estimators, each of whose sparsity patterns may be different.

Given the favorable theoretical properties of CQR under fixed dimensions, we expect the sparse CQR to also enjoy very competitive performance under the ultrahigh dimensional setting. However, there are few results in the literature to firmly establish such a claim, despite the massive literature on the sparse LS under ultrahigh dimensions. This is mainly caused by the severe nonsmoothness of the composite quantile loss. For example, even with a single quantile, the analysis of the lasso penalized QR was only recently done in [8], and the analysis therein is technically very different from the standard analysis for the least squares lasso. CQR uses the sum of many different quantile losses and hence makes it even more challenging to handle than the single QR estimator. The highly nonsmooth nature of the composite quantile loss is also a major obstacle for using standard algorithms for the penalized LS as its numeric solvers. In fact, coordinate descent, the most popular algorithm thus far for solving the least squares lasso, is not suitable for optimization problems involving a nonsmooth loss.

The contributions of this article are as follows. Firstly, we provide nonasymptotic analysis of both lasso and folded concave penalized CQR. Our analysis holds for very general fixed pair (n,p). As (n,p) go to infinity, we prove that the lasso estimator is estimation consistent under ultrahigh dimensions. Moreover, we show that the lasso estimator is tuning free, meaning that the rate of convergence is achieved by using an explicit penalization parameter. Secondly, we establish the oracle property for the folded concave penalized CQR and construct a new information criterion for calibrating the tuning parameter therein to give consistent model selection. Our paper

demonstrates a unified treatment of the concentration inequalities involving the CQR loss. Those inequalities could be of independent interest to studies on other models involving the check loss. Lastly, we exploit the structure of the composite quantile loss and design a specialized ADMM algorithm for efficiently computing the sparse penalized CQR estimator. The results in this work make sparse CQR a real alternative to the sparse LS for real applications.

The rest of the article is organized as follows. In Section II, we introduce the framework for the penalized CQR, followed by a discussion of the theoretical properties of the lasso and folded concave penalized CQR in Section III. We propose a new information criterion for selecting the tuning parameter and investigate its selection consistency in Section IV. In Section V, we present the efficient algorithm to solve the penalized CQR. Numerical studies are conducted in Section VI to show the superior finite-sample performance of penalized CQR over penalized LS. All proofs are relegated to Section VIII.

II. Penalized Composite Quantile Regression

Consider variable selection and coefficient estimation in the linear model

$$y = \beta_0 + \sum_{j=1}^{p} x_j \beta_j + \varepsilon, \tag{1}$$

where ε is independent of $\mathbf{x} = (x_1, \ldots, x_p)^{\mathrm{T}}$. Suppose β_0^* and $\beta^* = (\beta_1^*, \ldots, \beta_p^*)^{\mathrm{T}}$ are the true coefficients in model (1) that generate our i.i.d. data $(\mathbf{x}_i, y_i)_{i=1}^n$, where $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^{\mathrm{T}}$. Denote the response vector by $\mathbf{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$ and the design matrix by $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^{\mathrm{T}}$. We also write $\mathbf{X} = (X_1, \ldots, X_p)$, where $X_j = (x_{1j}, \ldots, x_{nj})^{\mathrm{T}}, 1 \le j \le p$. Let $\mathbb{X} = (X_0, \mathbf{X})$ be the augmented design with $X_0 = \mathbf{1}_n$ (corresponding to an intercept term), where $\mathbf{1}_n$ stands for the n-dimensional vector of all ones.

As mentioned in Section I, we consider the CQR rather than the LS or QR to estimate β in model (1). Assume that the random error ε has cumulative distribution function $F(\cdot)$ and probability density function $f(\cdot)$. To ensure identifiability of β_0, assume $F(0) = 1/2$. Given an ordered sequence of quantile levels $\tau_1 < \tau_2 < \cdots < \tau_K \in (0,1)$, let $\alpha_k^* = \beta_0^* + F^{-1}(\tau_k)$, where $F^{-1}(\tau_k) = \inf\{x : F(x) \ge \tau_k\}$ denotes the τ_k-th quantile of ε, $1 \le k \le K$. The canonical composite quantile regression estimates β by minimizing

$$\sum_{k=1}^{K} \sum_{i=1}^{n} \rho_{\tau_k}(y_i - \alpha_k - \mathbf{x}_i^{\mathrm{T}} \beta) \tag{2}$$

jointly over $\alpha = (\alpha_1, \ldots, \alpha_K) \in \mathbb{R}^K$ and $\beta \in \mathbb{R}^p$, where $\rho_{\tau_k}(u) = \{\tau_k - I(u < 0)\}u$ denotes the check loss for $1 \le k \le K$. A typical choice is to take equally spaced τ_k's: $\tau_k = k/(K+1), 1 \le k \le K$. As $K \to \infty$, [4] showed that the asymptotic efficiency of the CQR estimator relative to the LS estimator has a universal lower bound, $12\mathrm{var}(\varepsilon)\{\mathbb{E}_\varepsilon f(\varepsilon)\}^2$, which is at least 70.26% for an arbitrary error distribution and can be made arbitrarily large for non-normal distributions. The relative efficiency lower bound 0.7026 is further improved to 0.864 in [9]. Substantial efficiency gain can be achieved already with a relatively small K such as $K = 9$ or 19.

In the high-dimensional regime, the number of parameters p is typically large and may even exceed the number of observations n. Under the sparsity assumption on the model, assume that many components of β^* are zero. Let

$\mathscr{A} = \{1 \le j \le p\colon \beta_j^* \ne 0\}$ be the active set of β^* and denote the effective dimensionality of the model by $s = |\mathscr{A}|$. To harness the sparsity structure of β^*, let us consider the sparse penalized CQR

$$\min_{\alpha,\beta} \frac{1}{nK} \sum_{k=1}^{K} \sum_{i=1}^{n} \rho_{\tau_k}(y_i - \alpha_k - \mathbf{x}_i^{\mathrm{T}}\beta) + P_\lambda(\beta), \tag{3}$$

where $P_\lambda(\cdot)$ is a penalty function with regularization parameter λ. For instance, $P_\lambda(\cdot)$ can be the lasso [10], SCAD [11], MCP [12], and so on. In [4], the adaptive lasso was used to show the oracle property of the corresponding penalized estimator under fixed dimensions. The techniques used therein cannot be used to handle the ultrahigh dimensionality setting. As for the computation, the adaptive lasso penalized CQR was formulated as a linear program and was solved by a standard linear programming solver. However, such an approach does not scale well with dimensions. Other efficient alternatives are needed when p is large.

III. Analysis of Penalized Composite Quantile Regression

In this section, we show the theoretical properties of the penalized CQR under ultrahigh dimensionality with both lasso and folded concave penalties. All our results are nonasymptotic and holds for general (n, p). From these results, we establish the rate of convergence of the lasso penalized CQR and show its tuning free property. We also estabish the strong oracle property for a feasible solution of the folded concave penalized CQR by incorporating the lasso estimator as initial estimation. For ease of exposition, we introduce the following notation.

For $u \in \mathbb{R}$, let $u_+ = uI(u > 0)$ and $u_- = -uI(u < 0)$ be the positive and negative parts of u, respectively. Moreover, let $\mathrm{sgn}(u) = I(u > 0) - I(u < 0)$ be the sign function. The largest and smallest eigenvalues of a symmetric matrix $\mathbf{A}$ are denoted by $\Lambda_{\max}(\mathbf{A})$ and $\Lambda_{\min}(\mathbf{A})$, respectively. We also let ∂g be the subdifferential of a convex function g. For two matrices $\mathbf{A}_1, \mathbf{A}_2 \in \mathbb{R}^{m\times n}$, let $\langle \mathbf{A}_1, \mathbf{A}_2\rangle = \mathrm{tr}(\mathbf{A}_1^{\mathrm{T}}\mathbf{A}_2)$ be their trace inner product and $\|\mathbf{A}_1\|_F = \langle \mathbf{A}_1, \mathbf{A}_1\rangle^{1/2}$ be the Frobenius norm of $\mathbf{A}_1$. For any vector $\mathbf{v} = (v_1, \ldots, v_p)^{\mathrm{T}} \in \mathbb{R}^p$ and an arbitrary index set $I \subset \{1, \ldots, p\}$, we write $\mathbf{v}_I = (v_j, j \in I)^{\mathrm{T}}$ and denote by $\mathbf{X}_I = (X_j, j \in I)$ the submatrix consisting of the columns of $\mathbf{X}$ whose indices are in I. The complement of I is denoted by $I^{\mathrm{c}} = \{1, \ldots, p\}\backslash I$. For $q \in [1, \infty]$, the L_q-norm of $\mathbf{v}$ is denoted by $\|\mathbf{v}\|_q$.

A. Lasso penalized composite quantile regression

For $\lambda > 0$, we define the lasso penalized CQR estimator as

$$(\widehat{\alpha}_\lambda, \widehat{\beta}_\lambda) \coloneqq \arg\min_{\alpha,\beta} Q_n(\alpha, \beta) + \lambda \sum_{j=1}^{p} |\beta_j|, \tag{4}$$

where $Q_n(\alpha, \beta) = (nK)^{-1} \sum_{k=1}^{K} \sum_{i=1}^{n} \rho_{\tau_k}(y_i - \alpha_k - \mathbf{x}_i^{\mathrm{T}}\beta)$. In the sequel, we refer to $(\widehat{\alpha}_\lambda, \widehat{\beta}_\lambda)$ as the CQR lasso estimator.

For $\Delta \in \mathbb{R}^p$ and integer $m \ge 0$, let $\overline{\mathscr{A}}(\Delta, m) \subset \mathscr{A}^{\mathrm{c}}$ be the support of the m largest in absolute value components of $\Delta_{\mathscr{A}^{\mathrm{c}}}$. When $m = 0$, we take $\overline{\mathscr{A}}(\Delta, m)$ to be the empty set. The following assumption is imposed on the data and error distribution, which is typical in the QR literature.

(C0). The observations $(\mathbf{x}_i, y_i)_{i=1}^n$ are i.i.d. with $\min(n, p) \ge 3$. The density function is continuously differentiable and satisfies $f(u) \le \bar{f} < \infty$ and $f'(u) \le \bar{f}' \in (0, \infty)$ for all u in the support of ε. Moreover, there exists a

constant $\mathscr{U}_0 > 0$ such that $f(\alpha_k^* + u) \geq \underline{f} > 0$ for all $1 \leq k \leq K$ and $|u| \leq \mathscr{U}_0$. Also, $(\mathbf{x}_{i\mathscr{A}}, y_i)_{i=1}^n$ are in general positions (Section 2.2, [3]) and there is at least one continuous covariate in the true model.

Note that we do not impose any moment or light tail assumptions on the error distribution and the assumptions on the error density are mild and can be satisfied by many commonly seen distributions, including heavy-tailed distributions like Cauchy. We also assume that $\bar{f}$, $\bar{f}'$ and $\underline{f}$ are all positive constants. The assumptions on $(\mathbf{x}_{i\mathscr{A}}, y_i)$'s ensure that the CQR oracle estimator (5) is unique. This is a fairly common assumption in the QR literature (see [3], [13]). More discussions of the CQR oracle estimator can be found in Section III-B and Section B of the appendix.

We assume two additional conditions to establish the estimation consistency of the CQR lasso estimator, . For the sake of brevity, only fixed design is considered. Define the restricted set $\mathscr{C} = \big\{(\delta, \Delta) \in \mathbb{R}^K \times \mathbb{R}^p : \|\Delta_{\mathscr{A}^c}\|_1 \leq 3\|\Delta_{\mathscr{A}}\|_1 + \frac{3}{K}\|\delta\|_1\big\}$. The two assumptions are both imposed on the design matrix.

(C1). The design matrix $\mathbf{X}$ satisfies

$$\kappa_m = \inf_{(\delta,\Delta)\in\mathscr{C}, (\delta,\Delta)\neq\mathbf{0}} \frac{\sum_{k=1}^K \sum_{i=1}^n (\delta_k + \mathbf{x}_i^{\mathrm{T}}\Delta)^2}{n(K\|\Delta_{\mathscr{A}\cup\overline{\mathscr{A}}(\Delta,m)}\|_2^2 + \|\delta\|_2^2)} > 0.$$

(C2). The design matrix $\mathbf{X}$ satisfies

$$q = \frac{3}{8}\frac{\underline{f}^{3/2}}{\bar{f}'} \inf_{(\delta,\Delta)\in\mathscr{C}, (\delta,\Delta)\neq\mathbf{0}} \frac{\big[n^{-1}\sum_{i=1}^n \sum_{k=1}^K (\delta_k + \mathbf{x}_i^{\mathrm{T}}\Delta)^2\big]^{3/2}}{n^{-1}\sum_{i=1}^n \sum_{k=1}^K |\delta_k + \mathbf{x}_i^{\mathrm{T}}\Delta|^3} > 0.$$

Condition (C1) is an extension of the restricted identifiability property (RIP), also known as the restricted eigenvalue (RE) condition, to the case of the penalized CQR. RIP is a common assumption in the literature for sparse penalized regressions. For example, it is assumed in the penalized LS, Dantzig selector [14], [15] and penalized QR [8]. Condition (C2) is similar to the restricted nonlinearity assumption in [8]. The quantity q, referred to as the restricted nonlinear impact (RNI) coefficient by those authors, describes how well the CQR empirical loss function can be minorized by a quadratic function over the restricted set $\mathscr{C}$. We present in the following theorem the L_2-risk bound for the CQR lasso estimator, from which the estimation consistency of the estimator follows.

Theorem 1. *Under conditions (C0), (C1) and (C2), with probability at least* $1 - p_1(\lambda)$, *where*

$$p_1(\lambda) = 2K\exp\Big(-\frac{9n\lambda^2}{2}\Big) + 2p\exp\Big(-\frac{n\lambda^2}{2M_0}\Big) + \exp\bigg\{-2M_0\frac{s(1+\log p)}{\kappa_0}\bigg\},$$

the CQR lasso estimator $(\widehat{\alpha}_\lambda, \widehat{\beta}_\lambda)$ *satisfies*

$$\|\widehat{\alpha}_\lambda - \alpha^*\|_2 \leq \frac{8}{\underline{f}}\sqrt{\frac{K}{\kappa_m}}\bigg\{32\sqrt{\frac{2M_0}{\kappa_0}}\sqrt{\frac{1+\log p}{n}}(\sqrt{s}+1) + \lambda\sqrt{\frac{s}{\kappa_0}}\bigg\}$$

and for integer $m > 0$,

$$\|\widehat{\beta}_\lambda - \beta^*\|_2 \leq \frac{8}{\underline{f}\sqrt{\kappa_m}}\sqrt{1 + \frac{18s}{m} + \frac{18}{m}} \cdot\bigg\{32\sqrt{\frac{2M_0}{\kappa_0}}\sqrt{\frac{1+\log p}{n}}(\sqrt{s}+1) + \lambda\sqrt{\frac{s}{\kappa_0}}\bigg\},$$

provided that the growth condition

$$64\sqrt{\frac{2M_0}{\kappa_0}}\sqrt{\frac{1+\log p}{n}}(\sqrt{s}+1) + 2\lambda\sqrt{\frac{s}{\kappa_0}} \leq q\sqrt{\frac{\underline{f}}{K}}$$

holds, where $M_0 = \max_{0\le j\le p} \|X_j\|_2^2/n$.

Remark 1. By Theorem 1, one can typically choose the tuning parameter $\lambda = C\sqrt{\log p/n}$ for the CQR lasso estimator, where $C > \sqrt{2M_0}$ is some constant. For example, one can choose $C = 2\sqrt{M_0}$. Note that given the design $\mathbf{X}$, M_0 can be readily obtained. Therefore, in principle, the parameter λ in the lasso penalized CQR is tuning free. This is in similar spirit to the square-root lasso [16]. With such choice of λ, we can see that $p_1(\lambda) = \mathscr{O}(1)$ as $n, p \to \infty$, which leads to

$$\|\widehat{\beta}_\lambda - \beta^*\|_2 = \mathscr{O}_P\left(\frac{1}{\sqrt{\kappa_0\kappa_s}}\sqrt{\frac{s\log p}{n}}\right)$$

provided $q^{-1}\sqrt{s\log p/(n\kappa_0)} = \mathscr{o}(1)$ and $\kappa_0(s\log p)^{-1} = \mathscr{o}(1)$, by taking $m = s$. When κ_0 and κ_s are both positive constants, the CQR lasso estimator achieves the near-optimal rate $\sqrt{s\log p/n}$, which implies that it is a consistent estimator even when p is of exponential order of n, i.e., $\log p = \mathscr{O}(n^\gamma)$ for some $0 < \gamma < 1$, provided $s\log p = \mathscr{o}(n)$.

B. Folded concave penalized composite quantile regression

Folded concave penalized regression has been widely adopted in the statistical analysis of high-dimensional data due to its strong oracle optimality [17], [18]. In order to establish the oracle property of the folded concave penalized CQR estimator, let us first define the CQR oracle estimator,

$$(\widehat{\alpha}^{\mathrm{o}}, \widehat{\beta}^{\mathrm{o}}) := \underset{\alpha,\beta:\, \beta_{\mathscr{A}^{\mathrm{c}}}=\mathbf{0}}{\arg\min} \sum_{i=1}^{n}\sum_{k=1}^{K} \rho_{\tau_k}(y_i - \alpha_k - \mathbf{x}_i^{\mathrm{T}}\beta). \tag{5}$$

The oracle estimator $(\widehat{\alpha}^{\mathrm{o}}, \widehat{\beta}^{\mathrm{o}})$ is the ideal estimator one could possibly get using the CQR. It is not feasible in practice since $\mathscr{A}$ is unknown, but it serves as a benchmark estimator to which one can compare a penalized CQR estimator. In the following lemma, we show the rate of convergence of the CQR oracle estimator under the growing-dimension regime, i.e., the true dimensionality s is allowed to grow with n.

Let $\mathscr{A}_0 = \{0\} \cup \mathscr{A}$ and $\mathbb{X}_{\mathscr{A}_0} = (\mathbf{1}_n, \mathbf{X}_{\mathscr{A}})$. Denote $\underline{\mu} = \Lambda_{\min}(n^{-1}\mathbb{X}_{\mathscr{A}_0}^{\mathrm{T}}\mathbb{X}_{\mathscr{A}_0})$ and $\bar{\mu} = \Lambda_{\max}(n^{-1}\mathbb{X}_{\mathscr{A}_0}^{\mathrm{T}}\mathbb{X}_{\mathscr{A}_0})$. Moreover, let $M_{\mathscr{A}} = \max_{1\le i\le n}(s+1)^{-1}\big(1 + \|\mathbf{x}_{i\mathscr{A}}\|_2^2\big)$ and $M_{\mathscr{A}^{\mathrm{c}}} = \max_{1\le i\le n,\, j\in\mathscr{A}^{\mathrm{c}}} |x_{ij}|$. In this article, we assume that $M_{\mathscr{A}}$ and $M_{\mathscr{A}^{\mathrm{c}}}$ are both positive constants.

Lemma 1. *Under condition (C0), if* $32K(s+1)M_{\mathscr{A}}/(\sqrt{n}\underline{f}\,\underline{\mu}) \le \mathscr{U}_0$, *with probability at least* $1 - \exp\big[-(s+1)M_{\mathscr{A}}/(2\bar{\mu})\big]$, *the CQR oracle estimator satisfies*

$$\|\widehat{\alpha}^{o} - \alpha^*\|_2^2 + \|\widehat{\beta}^{o} - \beta^*\|_2^2 \le \frac{1024K^2(s+1)M_{\mathscr{A}}}{n\underline{f}^2\underline{\mu}^2}.$$

Remark 2. Assuming $s/(\underline{\mu}\sqrt{n}) = \mathscr{o}(1)$ and $\underline{\mu}/s = \mathscr{o}(1)$, the CQR oracle estimator has the following rate of convergence

$$\|\widehat{\alpha}^{\mathrm{o}} - \alpha^*\|_2 = \mathscr{O}_P\left(\frac{1}{\underline{\mu}}\sqrt{\frac{s}{n}}\right),\ \|\widehat{\beta}^{\mathrm{o}} - \beta^*\|_2 = \mathscr{O}_P\left(\frac{1}{\underline{\mu}}\sqrt{\frac{s}{n}}\right)$$

as $n \to \infty$. This implies that $\widehat{\beta}^{\mathrm{o}}$ is $\sqrt{n/s}$-consistent when s diverges with n, if we assume that $\underline{\mu} > 0$ is a fixed constant. Then it is required that $s = \mathscr{O}(n^\gamma)$ for some $0 < \gamma < \frac{1}{2}$.

Next, we introduce the details of the folded concave penalized CQR and show that the CQR oracle estimator is attainable via the folded concave penalized CQR. The folded concave penalized CQR at penalty level $\lambda > 0$ solves the following minimization problem

$$\min_{\alpha,\beta} Q_n(\alpha,\beta) + \sum_{j=1}^{p} p_\lambda(|\beta_j|), \tag{6}$$

where $p_\lambda(t), t \geq 0$ belongs to a class of folded concave penalties that satisfy the following properties:

(P1) $p_\lambda(t)$ is nondecreasing and concave in $t \geq 0$ and $p_\lambda(0) = 0$;

(P2) $p_\lambda(t)$ is differentiable in $t > 0$;

(P3) $p'_\lambda(t) \geq a_1\lambda, 0 < t \leq a_2\lambda$ and $p'_\lambda(0) := p'_\lambda(0+) \geq a_1\lambda$, where $a_1, a_2 > 0$ are fixed constants;

(P4) $p'_\lambda(t) = 0, t \geq a\lambda$ for a fixed constant $a > a_2$.

It can been shown that both the SCAD penalty and MCP belong to this class (see, e.g., [19]). For the analysis of the minimizer, we consider the local linear approximation (LLA, [20]) algorithm, where the initial estimator is chosen to be the CQR lasso estimator.

1) Initialize α and β with respectively $\widehat{\alpha}^{(0)}$ and $\widehat{\beta}^{(0)}$. Compute weights

$$\hat{w}_j^{(0)} = p'_\lambda(|\hat{\beta}_j^{(0)}|), j = 1, \ldots, p.$$

2) For $m = 1, 2, \ldots,$ repeat the LLA iterations in (2.a) and (2.b).

(2.a) Solve the following convex optimization problem for $\widehat{\alpha}^{(m)}$ and $\widehat{\beta}^{(m)}$

$$(\widehat{\alpha}^{(m)}, \widehat{\beta}^{(m)}) := \arg\min_{\alpha,\beta} \; Q_n(\alpha,\beta) + \sum_{j=1}^{p} \hat{w}_j^{(m-1)} |\beta_j|.$$

(2.b) Calculate the weights

$$\hat{w}_j^{(m)} = p'_\lambda(|\hat{\beta}_j^{(m)}|), \; j = 1, \ldots, p.$$

In order to establish the oracle property, we assume a "beta-min" condition, i.e., the true coefficient exhibit sufficient signal:

(C3) $\min_{j \in \mathscr{A}} |\beta_j^*| > (a+1)\lambda$.

The "beta-min" condition is always assumed for non-convexly penalized regressions and is almost a necessary condition for establishing consistency results. See, e.g., [11], [13], [18].

Theorem 2. *Suppose the folded concave penalized CQR* (6) *is solved with the LLA algorithm that is initialized with the CQR lasso estimator* (4) *at penalty level* $\lambda_0 = c\sqrt{\log p/n}$ *for some constant* $c > \sqrt{2M_0}$. *Let* $r_0 = \min_{j \in \mathscr{A}} |\beta_j^*| - a\lambda$ *and* $r_* = \sqrt{(s+1)M_{\mathscr{A}} \log n/n}$. *Assume the folded concave penalty* $p_\lambda(\cdot)$ *satisfies properties (P1) – (P4), where for integer* $m > 0$, λ *is taken such that*

$$\begin{aligned} \lambda \geq & \frac{8}{a_0 \underline{f} \sqrt{\kappa_m}} \sqrt{1 + \frac{18s}{m} + \frac{18}{m}} \\ & \cdot \left\{ 32 \sqrt{\frac{2M_0}{\kappa_0}} \sqrt{\frac{1+\log p}{n}} (\sqrt{s}+1) + \lambda_0 \sqrt{\frac{s}{\kappa_0}} \right\}. \end{aligned} \tag{7}$$

Under conditions (C0) – (C3) and the assumptions that $r_0\sqrt{(s+1)M_{\mathscr{A}}} \le \mathscr{U}_0$, $r_*\sqrt{(s+1)M_{\mathscr{A}}} \le \mathscr{U}_0$ *and* $\lambda > 8K(\underline{f}\underline{\mu})^{-1}\sqrt{(s+1)M_{\mathscr{A}}/n}$, *with probability at least* $p_0 = 1 - p_1(\lambda_0) - p_2(r_0) - p_2(r_*) - p_3$, *the LLA algorithm converges to the oracle estimator* $(\widehat{\alpha}^o, \widehat{\beta}^o)$ *in two iterations, where* $p_1(\cdot)$ *is given in Theorem 1,* $p_2(\cdot)$ *is defined as*

$$p_2(r) = \exp\left\{-\frac{n(t(r))^2}{32\bar{\mu}r^2}\right\}, \text{ where } t(r) = \frac{\underline{f}}{4K}\underline{\mu}r^2 - 2r\sqrt{\frac{(s+1)M_{\mathscr{A}}}{n}},$$

and

$$p_3 = 2(p-s)\exp\left(-\frac{2nB^2}{M_0}\right) + 2(p-s)n^{2(K+s)}\exp\left(-\frac{3nB^2}{24\bar{f}M_{\mathscr{A}^c}^2\bar{\mu}^{1/2}r_* + 8M_{\mathscr{A}^c}B}\right)$$
$$+ 2(p-s)n^{2(K+s)}\exp\left(-\frac{3nB_0^2}{24\bar{f}M_{\mathscr{A}^c}^2(s+1)^{1/2}M_{\mathscr{A}}^{1/2}n^{-2}r_* + 4M_{\mathscr{A}^c}B_0}\right)$$

with $B = \frac{1}{2}\left[a_1\lambda - \frac{K+s}{n}M_{\mathscr{A}^c} - \bar{f}M_{\mathscr{A}^c}\bar{\mu}^{1/2}r_*\right]_+$ *and* $B_0 = \left[\frac{B}{2} - \frac{8\bar{f}M_{\mathscr{A}^c}\sqrt{(s+1)M_{\mathscr{A}}}}{n^2}r_*\right]_+$.

It is easy to translate Theorem 2 into an asymptotic statement that the folded concave penalized CQR estimator finds the oracle CQR estimator with overwhelming probability. For brevity, we omit such discussions. We emphasize that unlike the lasso penalized CQR, the regularization parameter in the folded concave penalized CQR involves unknown quantities. In order to apply Theorem 2 in applications, we need a data-driven choice of the regularization parameter. To this end, we construct a new information criterion for selecting the tuning parameter in the next section.

IV. Tuning Parameter Calibration

For the folded concave penalized CQR, there exists a tuning sequence λ_n (we write λ_n here to signify its dependence on n) such that the LLA algorithm yields the CQR oracle estimator in two iterations with probability approaching one (Theorem 2). However, as we pointed out already, there is no direct way to use such λ_n as given in Theorem 2, since it relies on unknown quantities. We thus pursue data-driven approach to the selection of λ. Consider the following high-dimensional Bayesian information criterion:

$$\mathrm{BIC}^{\mathrm{H}}(\lambda) = \frac{1}{nK}\sum_{i=1}^{n}\sum_{k=1}^{K}\rho_{\tau_k}(y_i - \hat{\alpha}_k^{\lambda} - \mathbf{x}_i^{\mathrm{T}}\hat{\beta}^{\lambda}) + |\hat{A}_{\lambda}|\frac{C_n\log(p)}{n}, \tag{8}$$

where $(\widehat{\alpha}^{\lambda}, \widehat{\beta}^{\lambda})$ is the two-step estimator from the LLA algorithm initialized by the CQR lasso estimator with regularization parameter λ_0 (Theorem 2), $\hat{A}_{\lambda} = \{1 \le j \le p : \hat{\beta}_j^{\lambda} \ne 0\}$ and C_n is a positive number depending on n (allowed to grow with n). We compare the values of $\mathrm{BIC}^{\mathrm{H}}(\lambda)$ for $\lambda \in \Xi_n = \{\lambda : |\hat{A}_{\lambda}| \le J_n\}$, where $J_n > s$ represents a rough estimate of the upper bound of the model sparsity and is allowed to (slowly) diverge as $n \to \infty$. Typically, J_n is much smaller than p, so that one can avoid searching over a notoriously large model space. The tuning parameter selected via BIC is given by

$$\hat{\lambda}_n = \arg\min_{\lambda\in\Xi_n}\mathrm{BIC}^{\mathrm{H}}(\lambda).$$

We call the information criterion BIC because it can be shown in the following Theorem that $\Pr(\hat{A}_{\hat{\lambda}_n} = \mathscr{A}) \to 1$ as $n \to \infty$. Note that model selection consistency is the signature property of BIC in the fixed-dimension setting.

Let $M = \max_{1\le i\le n, 0\le j\le p} |x_{ij}|$ and assume M is a positive constant. Also, define

$$\underline{\zeta} = \inf_{A\supset\mathscr{A}, |A|\le 2J_n} \Lambda_{\min}(n^{-1}(\mathbf{1}_n, \mathbf{X}_A)^{\mathrm{T}}(\mathbf{1}_n, \mathbf{X}_A))$$

and

$$\bar{\zeta} = \sup_{A\supset\mathscr{A}, |A|\le 2J_n} \Lambda_{\max}(n^{-1}(\mathbf{1}_n, \mathbf{X}_A)^{\mathrm{T}}(\mathbf{1}_n, \mathbf{X}_A)),$$

and assume that both $\underline{\zeta}$ and $\bar{\zeta}$ are positive constants, and so are $\kappa_0, \kappa_s, \underline{\mu}$ and $\bar{\mu}$.

Theorem 3. *Under the conditions of Theorem 2, and as $n\to\infty$, assuming that $s = \mathscr{O}(1), n = o(p), J_n^2 \log p = o(n), C_n \to \infty$ and $\sqrt{\max(J_n, C_n)\log p/n} = o(\min_{j\in\mathscr{A}} |\beta_j|)$, the criterion $BIC^H(\lambda)$ is selection consistent, i.e., $\Pr(\hat{A}_{\hat{\lambda}_n} = \mathscr{A}) \to 1$ as $n\to\infty$.*

Remark 3. The sequence C_n is often taken to slowly diverge to infinity, e.g., $C_n = \log\log n$. Under fixed model sparsity, $s = \mathscr{O}(1)$, it is implied from Theorem 3 that $\mathrm{BIC}^{\mathrm{H}}(\lambda)$ is consistent when $\log p = \mathscr{O}(n^{\gamma_1})$ and $J_n = \mathscr{O}(n^{\gamma_2})$ for some positive constants γ_1 and γ_2 such that $\gamma_1 + 2\gamma_2 < 1$. It is worth mentioning that the fixed model sparsity is assumed in order to achieve ultrahigh dimensionality due to technical difficulties with the check loss (see, e.g., [21]). If instead we allow the model sparsity to grow, using current technique, we must assume p can be of at most polynomial order of n (see, e.g., [22]).

One problem with the criterion in (8) is that it is not scale invariant. In practice, one needs to standardize the variables beforehand. Therefore, we also consider a scale invariant version of the high-dimensional Bayesian information criterion:

$$\mathrm{BIC}^{\mathrm{HL}}(\lambda) = \log\left(\frac{1}{K}\sum_{i=1}^{n}\sum_{k=1}^{K}\rho_{\tau_k}(y_i - \hat{\alpha}_k^{\lambda} - \mathbf{x}_i^{\mathrm{T}}\hat{\beta}^{\lambda})\right) + |\hat{A}_{\lambda}|\frac{C_n\log(p)}{n}. \tag{9}$$

Theorem 4. *In model* (1)*, assume that $\mathbb{E}(|\varepsilon|) < \infty$. Under the conditions of Theorem 2, as $n\to\infty$, assume moreover that $s = \mathscr{O}(1), n = o(p), J_n^2\log p = o(n), C_n\to\infty$ and $\sqrt{\max(J_n, C_n)\log p/n} = o(\min_{j\in\mathscr{A}}|\beta_j|)$. The criterion $BIC^{HL}(\lambda)$ is selection consistent, i.e., $\Pr(\hat{A}_{\hat{\lambda}_n} = \mathscr{A}) \to 1$ as $n\to\infty$.*

Remark 4. The selection consistency of $\mathrm{BIC}^{\mathrm{HL}}(\lambda)$ requires additionally that $\mathbb{E}(|\varepsilon|) < \infty$. Our empirical study in Section VI suggests that this might be a necessary condition. Indeed, in the numerical comparison there, we see that $\mathrm{BIC}^{\mathrm{HL}}(\lambda)$ does not perform well under the Cauchy error.

V. Optimization

Note that both lasso and folded concave penalized CQR can be solved with one or more runs of the following weighted lasso penalized CQR:

$$\min_{\alpha,\beta}\frac{1}{nK}\sum_{k=1}^{K}\sum_{i=1}^{n}\rho_{\tau_k}(y_i - \alpha_k - \mathbf{x}_i^{\mathrm{T}}\beta) + \lambda_1\sum_{j=1}^{p} d_j|\beta_j|, \tag{10}$$

where $\lambda_1 > 0$ and $d_j \ge 0$ for $j = 1,\ldots,p$. Specifically, the CQR lasso estimator can be achieved by letting $d_j = 1$ for all $j = 1,\ldots,p$, while the folded concave penalized CQR estimator can be obtained by iteratively solving (10)

with $d_j = \hat{w}_j^{(m-1)}$ in the mth LLA iteration. Hence, in the sequel, we only need to focus on developing the algorithm for solving (10).

Traditionally, (10) can be solved by linear programming if n and p are moderate. However, linear programming does not scale well when p is large [23]. We hence propose an efficient alternating direction method of multipliers (ADMM) algorithm for solving (10). The algorithm is based on a reformulation that turns the original problem into one that can harness the power of ADMM. We point out that there are multiple ways to formulate (10) into problems that are solvable by ADMM. For instance, in a relevant context, two ADMM versions are proposed in [23] to efficiently solve the penalized QR and they can be readily modified to solve (10). However, the formulation we present here is different from the ideas in [23] and it results in a more stable ADMM algorithm. To elaborate on our algorithm, let $z_{ik} = y_i - \alpha_k - \mathbf{x}_i^{\mathrm{T}}\beta$ for $i = 1,\ldots,n$ and $k = 1,\ldots,K$, and define matrix $\mathbf{Z} = (z_{ik})_{n\times K}$ in terms of the z_{ik}'s. By convexity, it can be immediately seen that (10) is equivalent to

$$\begin{aligned} \text{minimize} \quad & \frac{1}{nK}\sum_{k=1}^{K}\sum_{i=1}^{n}\rho_{\tau_k}(z_{ik}) + \lambda_1\sum_{j=1}^{p} d_j|\gamma_j| \\ \text{subject to} \quad & \mathbf{Z} = \mathbf{1}_K^{\mathrm{T}}\otimes\mathbf{y} - \mathbf{1}_n\otimes\alpha^{\mathrm{T}} - \mathbf{1}_K^{\mathrm{T}}\otimes(\mathbf{X}\beta) \\ & \gamma = \beta, \end{aligned} \tag{11}$$

where $\gamma = (\gamma_1,\ldots,\gamma_p)^{\mathrm{T}}$ and $\otimes$ denotes the Kronecker product. As will be seen, the introduction of the γ_j's renders a more stable ADMM algorithm where only the dual updates involve the non-smooth functions. For ease of notation, let $\mathbf{Y} = \mathbf{1}_K\otimes\mathbf{y}$, $\varphi = (\alpha^{\mathrm{T}},\beta^{\mathrm{T}})^{\mathrm{T}}$,

$$\mathbb{X}_1 = \begin{pmatrix} \mathbf{1}_n & \cdots & \mathbf{0} & \mathbf{X} \\ \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \cdots & \mathbf{1}_n & \mathbf{X} \end{pmatrix}_{(nK)\times(p+K)}, \text{ and } \mathbb{X}_2 = (\mathbf{O}_{p\times K} \;\; \mathbf{I}_p)_{p\times(p+K)}.$$

Then, (11) can be equivalently written as

$$\begin{aligned} \text{minimize} \quad & \frac{1}{nK}\sum_{k=1}^{K}\sum_{i=1}^{n}\rho_{\tau_k}(z_{ik}) + \lambda_1\sum_{j=1}^{p} d_j|\gamma_j| \\ \text{subject to} \quad & \begin{pmatrix} \mathbb{X}_1 \\ -\mathbb{X}_2 \end{pmatrix}\varphi + \begin{pmatrix} \mathrm{vec}(\mathbf{Z}) \\ \gamma \end{pmatrix} = \begin{pmatrix} \mathbf{Y} \\ \mathbf{0} \end{pmatrix}, \end{aligned} \tag{12}$$

where vec stands for the vectorization operator that stacks the columns of a matrix one underneath the other to form a single vector. The augmented Lagrangian of problem (12) is

$$\begin{aligned} L_\sigma(\varphi,\mathbf{Z},\gamma,\mathbf{U},\mathbf{v}) := {} & \frac{1}{nK}\sum_{k=1}^{K}\sum_{i=1}^{n}\rho_{\tau_k}(z_{ik}) + \lambda_1\sum_{j=1}^{p} d_j|\gamma_j| \\ & + \langle \mathrm{vec}(\mathbf{U}), \mathrm{vec}(\mathbf{Z}) + \mathbb{X}_1\varphi - \mathbf{Y}\rangle + \langle \mathbf{v}, \gamma - \mathbb{X}_2\varphi\rangle \\ & + \frac{\sigma}{2}\|\mathrm{vec}(\mathbf{Z}) + \mathbb{X}_1\varphi - \mathbf{Y}\|_F^2 + \frac{\sigma}{2}\|\gamma - \mathbb{X}_2\varphi\|_2^2, \end{aligned} \tag{13}$$

where $\mathbf{U} = (u_{ik})_{n\times K}$ and $\mathbf{v} = (v_1,\ldots,v_p)^{\mathrm{T}}$ are the Lagrangian multipliers and $\sigma > 0$. Let $\varphi^r, \mathbf{Z}^r, \gamma^r, \mathbf{U}^r$ and $\mathbf{v}^r$ be the iterate after the rth iteration of the algorithm, where $r \geq 0$. The ADMM has the following updates in the $(r+1)$st

iteration

$$\begin{cases} \boldsymbol{\varphi}^{r+1} := \arg\min_{\boldsymbol{\varphi}} L_\sigma(\boldsymbol{\varphi}, \mathbf{Z}^r, \boldsymbol{\gamma}^r, \mathbf{U}^r, \mathbf{v}^r), \\ (\mathbf{Z}^{r+1}, \boldsymbol{\gamma}^{r+1}) := \arg\min_{\mathbf{Z},\boldsymbol{\gamma}} L_\sigma(\boldsymbol{\varphi}^{r+1}, \mathbf{Z}, \boldsymbol{\gamma}, \mathbf{U}^r, \mathbf{v}^r), \\ \mathrm{vec}(\mathbf{U}^{r+1}) := \mathrm{vec}(\mathbf{U}^r) + \sigma\{\mathrm{vec}(\mathbf{Z}^{r+1}) + \mathbb{X}_1\boldsymbol{\varphi}^{r+1} - \mathbf{Y}\}, \\ \mathbf{v}^{r+1} := \mathbf{v}^r + \sigma\{\boldsymbol{\gamma}^{r+1} - \mathbb{X}_2\boldsymbol{\varphi}^{r+1}\}. \end{cases} \tag{14}$$

It follows from (13) that

$$\boldsymbol{\varphi}^{r+1} = \frac{1}{\sigma}(\mathbb{X}_1^{\mathrm{T}}\mathbb{X}_1 + \mathbb{X}_2^{\mathrm{T}}\mathbb{X}_2)^{-1}\{\mathbb{X}_1^{\mathrm{T}}(\sigma\mathbf{Y} - \sigma\mathrm{vec}(\mathbf{Z}^r) - \mathrm{vec}(\mathbf{U}^r)) + \mathbb{X}_2^{\mathrm{T}}(\sigma\boldsymbol{\gamma}^r + \mathbf{v}^r)\}.$$

Note that

$$\mathbb{X}_1^{\mathrm{T}}\mathbb{X}_1 + \mathbb{X}_2^{\mathrm{T}}\mathbb{X}_2 = \begin{pmatrix} n\mathbf{I}_K & \mathbf{1}_K\mathbf{1}_n^{\mathrm{T}}\mathbf{X} \\ \mathbf{X}^{\mathrm{T}}\mathbf{1}_n\mathbf{1}_K^{\mathrm{T}} & \mathbf{I}_p + K\mathbf{X}^{\mathrm{T}}\mathbf{X} \end{pmatrix}.$$

Let the Schur complement of $n\mathbf{I}_K$ in the above matrix be

$$\mathbf{S} = \mathbf{I}_p + K\mathbf{X}^{\mathrm{T}}\mathbf{X} - \frac{1}{n}(\mathbf{X}^{\mathrm{T}}\mathbf{1}_n\mathbf{1}_K^{\mathrm{T}})(\mathbf{1}_K\mathbf{1}_n^{\mathrm{T}}\mathbf{X}) = \mathbf{I}_p + K\mathbf{X}_0^{\mathrm{T}}\mathbf{X}_0,$$

where $\mathbf{X}_0 = (\mathbf{I}_n - n^{-1}\mathbf{1}_n\mathbf{1}_n^{\mathrm{T}})\mathbf{X}$ is the centered design matrix. Then, we have

$$(\mathbb{X}_1^{\mathrm{T}}\mathbb{X}_1 + \mathbb{X}_2^{\mathrm{T}}\mathbb{X}_2)^{-1} = \begin{pmatrix} \frac{1}{n}\mathbf{I}_K + \frac{1}{n^2}\mathbf{1}_K\mathbf{1}_n^{\mathrm{T}}\mathbf{X}\mathbf{S}^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{1}_n\mathbf{1}_K^{\mathrm{T}} & -\frac{1}{n}\mathbf{1}_K\mathbf{1}_n^{\mathrm{T}}\mathbf{X}\mathbf{S}^{-1} \\ -\frac{1}{n}\mathbf{S}^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{1}_n\mathbf{1}_K^{\mathrm{T}} & \mathbf{S}^{-1} \end{pmatrix}.$$

When p is large, the computation of $\mathbf{S}^{-1}$ can be expensive. We can apply the Sherman–Morrison–Woodbury formula to get

$$\mathbf{S}^{-1} = \mathbf{I}_p - K\mathbf{X}_0^{\mathrm{T}}(\mathbf{I}_n + K\mathbf{X}_0\mathbf{X}_0^{\mathrm{T}})^{-1}\mathbf{X}_0,$$

where we only need to evaluate the inverse of an $n \times n$ matrix. When n is relatively small compared to p, this formula can be very helpful.

Remark 5. In the actual implementation, we often center the design matrix before fitting the model. Then, $\mathbb{X}_1^{\mathrm{T}}\mathbb{X}_1 + \mathbb{X}_2^{\mathrm{T}}\mathbb{X}_2$ is block diagonal since $\mathbf{X}^{\mathrm{T}}\mathbf{1}_n = \mathbf{0}$ and its inverse can be readily obtained.

The update of $\mathbf{Z}^{r+1}$ and $\boldsymbol{\gamma}^{r+1}$ can be carried out component-wisely. This pertains to the application of the proximity operator of the check loss $\rho_\tau(\cdot)$ and the absolute value function $|\cdot|$, respectively. For $v \in \mathbb{R}$, the proximity operator of $\rho_\tau(\cdot)$ with respect to a parameter $a > 0$ is defined as

$$\mathrm{Prox}_{\rho_\tau}(v, a) := \operatorname*{arg\,min}_{u\in\mathbb{R}} \rho_\tau(u) + \frac{a}{2}(u - v)^2.$$

The following lemma gives the closed form expression of $\mathrm{Prox}_{\rho_\tau}$.

Lemma 2. *For $v \in \mathbb{R}$ and $a > 0$, the proximity operator of the check loss $\rho_\tau(\cdot)$ with respect to parameter a is given by*

$$Prox_{\rho_\tau}(v, a) = v - \max\Big(\frac{\tau - 1}{a}, \min\Big(v, \frac{\tau}{a}\Big)\Big).$$

Now by Lemma 2, we obtain

$$z_{ik}^{r+1} = \mathrm{Prox}_{\rho_{\tau_k}}\left(y_i - \alpha_k^{r+1} - \mathbf{x}_i^{\mathrm{T}}\beta^{r+1} - \frac{u_{ik}^r}{\sigma}, nK\sigma\right), 1 \le i \le n, 1 \le k \le K.$$

The proximity operator of $|\cdot|$ is the soft-thresholding operator and thus

$$\gamma_j^{r+1} = \mathrm{Shrink}\left(\beta_j^{r+1} - \frac{v_j^r}{\sigma}, \frac{\lambda_1 d_j}{\sigma}\right),$$

where $\mathrm{Shrink}(v,a) = \mathrm{sgn}(v)(|v|-a)_+$.

We summarize the above ADMM algorithm in Algorithm 1. A discussion of the convergence criterion for this algorithm can be found in the appendix.

Algorithm 1: The ADMM algorithm for solving the weighted lasso penalized composite quantile regression

1) Initialize the algorithm with $(\varphi^0, \mathbf{Z}^0, \gamma^0, \mathbf{U}^0, \mathbf{v}^0)$, where $\varphi^0 = ((\alpha^0)^{\mathrm{T}}, (\beta^0)^{\mathrm{T}})^{\mathrm{T}}$.
2) For $r = 0, 1, 2, \ldots,$ repeat steps (2.1) – (2.3) until convergence.

(2.1) Update

$$\varphi^{r+1} = ((\alpha^{r+1})^{\mathrm{T}}, (\beta^{r+1})^{\mathrm{T}})^{\mathrm{T}} \leftarrow \frac{1}{\sigma}(\mathbb{X}_1^{\mathrm{T}}\mathbb{X}_1 + \mathbb{X}_2^{\mathrm{T}}\mathbb{X}_2)^{-1} \cdot \{\mathbb{X}_1^{\mathrm{T}}(\sigma\mathbf{Y} - \sigma\mathrm{vec}(\mathbf{Z}^r) - \mathrm{vec}(\mathbf{U}^r)) + \mathbb{X}_2^{\mathrm{T}}(\sigma\gamma^r + \mathbf{v}^r)\}.$$

(2.2) Update

$$z_{ik}^{r+1} \leftarrow \mathrm{Prox}_{\rho_{\tau_k}}\left(y_i - \alpha_k^{r+1} - \mathbf{x}_i^{\mathrm{T}}\beta^{r+1} - \frac{u_{ik}^r}{\sigma}, nK\sigma\right), 1 \le i \le n, 1 \le k \le K,$$

and

$$\gamma_j^{r+1} \leftarrow \mathrm{Shrink}\left(\beta_j^{r+1} - \frac{v_j^r}{\sigma}, \frac{\lambda_1 d_j}{\sigma}\right), 1 \le j \le p.$$

(2.3) Update

$$\mathrm{vec}(\mathbf{U}^{r+1}) \leftarrow \mathrm{vec}(\mathbf{U}^r) + \sigma\{\mathrm{vec}(\mathbf{Z}^{r+1}) + \mathbb{X}_1\varphi^{r+1} - \mathbf{Y}\}$$

and

$$\mathbf{v}^{r+1} \leftarrow \mathbf{v}^r + \sigma\{\gamma^{r+1} - \mathbb{X}_2\varphi^{r+1}\}.$$

VI. Numerical Experiments

We conduct Monte Carlo studies to assess the finite sample performance of the proposed method as well as the tuning criterion. First, we compare the estimators from the penalized LS, the penalized CQR, the ideal oracle LS, and the oracle CQR. Recall that the oracle estimators are obtained through applying the canonical LS and CQR to the true underlying model. Second, we compare the tuned penalized CQR estimation by using cross-validation (CV) and by using the proposed information criteria.

Our simulated data are from the linear model

$$y = \beta_0^* + \mathbf{x}^{\mathrm{T}}\beta^* + \varepsilon, \tag{15}$$

where $\beta_0^* = 0$ and $\beta^* = (3, 1.5, 0, 0, 2, \mathbf{0}_{p-5})^{\mathrm{T}}$. The covariates are drawn from the multivariate normal distribution, $\mathbf{x} \sim N_p(\mathbf{0}, \Sigma)$, where two different covariance matrices $\Sigma = (0.5^{|i-j|})$ and $\Sigma = (0.8^{|i-j|})$ are considered. For the error distribution, we refer to [4] and consider five different shapes:

(a) the normal distribution, $\varepsilon \sim N(0,3)$;

(b) the mixture normal distribution (MN), $\varepsilon \sim \sqrt{6} \times \varepsilon^*$, where $\varepsilon^* \sim 0.5N(0,1) + 0.5N(0,0.5^6)$;

(c) the mixture double gamma distribution (MDG), $\varepsilon \sim \frac{1}{9}\varepsilon^*$, where $\varepsilon^* \sim f(\varepsilon) = e^{-14} \cdot \frac{1}{2}e^{-|\varepsilon|} + (1-e^{-14}) \cdot \frac{1}{\Gamma(15)}|\varepsilon|^{14}e^{-|\varepsilon|}$;

(d) the t-distribution with 3 degrees of freedom $\varepsilon \sim t_3$; and,

(e) the Cauchy distribution, $\varepsilon \sim$ Cauchy.

In the simulation study, our training data are composed of n observations $(\mathbf{x}_i, y_i)_{i=1}^n$, independently generated from model (15). An independent set of n observations is also simulated from the same model for parameter tuning of the training model. We evaluate the variable selection performance of the estimated coefficients $\widehat{\beta}$ by the number of false positives $\text{FP} = |\hat{A} \backslash A^*|$ and the number of false negatives $\text{FN} = |A^* \backslash \hat{A}|$, where $A^* = \{1 \le j \le p \colon \beta_j^* \neq 0\}$ and $\hat{A} = \{1 \le j \le p \colon \hat{\beta}_j \neq 0\}$. The estimation accuracy of $\widehat{\beta}$ is measured by the model error $(\widehat{\beta} - \beta^*)^{\mathrm{T}} \Sigma (\widehat{\beta} - \beta^*)$. Two sets of data dimensions $(n,p) = (100,600)$ and $(n,p) = (200,1200)$ are used in our simulations. In all settings, we use $K = 19$ quantile levels $\tau_k = 0.05k$, $k = 1, \ldots, 19$. The simulation results are summarized in Tables I and II.

It can be seen from the tables that the CQR oracle estimator has very similar model error to the LS estimator under normal error, while is more efficient under the other error distributions. In particular, the model error of the LS estimator explodes under the Cauchy error. In theory, it can be arbitrarily large. SCAD penalized CQR estimators have very close model errors to the CQR oracle estimator under most error distributions and outperform the penalized LS estimators. In terms of model selection accuracy, the SCAD penalized CQR estimator also outperform all the other penalized estimators.

The comparison between CV, BIC^{H} and BIC^{HL} for tuning parameter selection is shown in Tables III and IV. Note that BIC^{HL} does not perform well under the Cauchy error. This confirms its requirement for the first moment of the error distribution. The information criterion is computationally more efficient than CV and also delivers better results.

VII. Discussion

In this article, we studied the sparse penalized CQR under various forms of regularization. In particular, we established the estimation consistency of the CQR lasso estimator. Through the LLA algorithm, we showed that the CQR oracle estimator could be achieved via folded concave penalized CQR. Our theoretical analysis remains valid even when the dimensionality is ultrahigh in the sense that $p = \mathscr{O}(n^{\nu})$ with $0 < \nu < \frac{1}{2}$.

We also developed a fast and stable ADMM algorithm for solving the weighted L_1-penalized CQR. Numerical studies proved the efficiency of the algorithm. The methodologies and numerical solvers proposed in this article make the sparse CQR a real alternative to the sparse LS in practice. It can be applied whenever the estimation efficiency of the coefficients are concerned.

TABLE I

SIMULATION RESULTS FOR THE NUMERICAL COMPARISON OF FOUR METHODS: LS-LASSO, LS-SCAD, CQR-LASSO AND CQR-SCAD, UNDER MODEL (15) WITH $n = 100$ AND $p = 600$. THE LS-ORACLE AND CQR-ORACLE SERVE AS THE BENCHMARK. TWO COVARIANCE STRUCTURES $\Sigma = (0.5^{|i-j|})$ AND $\Sigma = (0.8^{|i-j|})$ ARE SHOWN, UNDER EACH OF WHICH FIVE ERROR DISTRIBUTIONS ARE CONSIDERED: $N(0,3)$, MIXTURE NORMAL, MIXTURE DOUBLE GAMMA, t_3 AND CAUCHY. THE ESTIMATION ACCURACY IS REPORTED AS "MODEL ERROR" AND THE SELECTION ACCURACY IS REPORTED AS "FP, FN". NUMBERS LISTED ARE AVERAGES OVER 100 INDEPENDENT RUNS, WITH STANDARD ERRORS REPORTED IN THE PARENTHESES

		$N(0,3)$	MN	MDG	t_3	Cauchy
			$\Sigma = (0.5^{\|i-j\|})$			
Model error	LS-oracle	0.093	0.093	0.084	0.098	9350.072
		(0.008)	(0.007)	(0.007)	(0.009)	(6837.507)
	CQR-oracle	0.105	0.004	0.025	0.047	0.094
		(0.008)	(0.002)	(0.003)	(0.004)	(0.011)
	LS-lasso	0.664	0.620	0.588	0.663	18.963
		(0.035)	(0.025)	(0.031)	(0.054)	(1.513)
	LS-SCAD	0.671	0.646	0.523	0.578	31.738
		(0.038)	(0.036)	(0.033)	(0.036)	(7.959)
	CQR-lasso	0.792	0.272	0.465	0.374	1.672
		(0.041)	(0.029)	(0.034)	(0.022)	(0.144)
	CQR-SCAD	0.122	0.006	0.032	0.064	0.438
		(0.019)	(0.002)	(0.004)	(0.006)	(0.098)
FP, FN	LS-lasso	16.55, 0	16.91, 0	16.53, 0	15.83, 0	13.37, 1.79
		(1.28), (0)	(0.92), (0)	(1.07), (0)	(1.08), (0)	(2.41), (0.12)
	LS-SCAD	18.04, 0	18.00, 0	17.04, 0	16.81, 0	17.11, 1.78
		(1.54), (0)	(1.44), (0)	(1.58), (0)	(1.10), (0)	(2.93), (0.13)
	CQR-lasso	15.33, 0	15.29, 0	14.49, 0	12.89, 0	38.75, 0.01
		(0.67), (0)	(0.67), (0)	(0.53), (0)	(0.55), (0)	(2.93), (0.01)
	CQR-SCAD	1.68, 0.01	1.62, 0	2.27, 0	2.33, 0	2.18, 0.01
		(0.25), (0.01)	(0.29), (0)	(0.32), (0)	(0.34), (0)	(0.38), (0.01)
			$\Sigma = (0.8^{\|i-j\|})$			
Model error	LS-oracle	0.097	0.092	0.079	0.088	184.788
		(0.008)	(0.008)	(0.006)	(0.008)	(91.476)
	CQR-oracle	0.097	0.005	0.023	0.046	0.134
		(0.008)	(0.002)	(0.002)	(0.004)	(0.011)
	LS-lasso	0.488	0.493	0.441	0.422	19.649
		(0.025)	(0.028)	(0.021)	(0.031)	(1.623)
	LS-SCAD	0.524	0.443	0.422	0.442	27.706
		(0.026)	(0.020)	(0.019)	(0.029)	(5.775)
	CQR-lasso	0.498	0.152	0.259	0.269	0.993
		(0.029)	(0.023)	(0.029)	(0.014)	(0.087)
	CQR-SCAD	0.123	0.005	0.032	0.060	0.355
		(0.013)	(0.001)	(0.003)	(0.005)	(0.055)
FP, FN	LS-lasso	13.06, 0	12.34, 0	11.97, 0	13.59, 0	9.21, 1.60
		(0.85), (0)	(0.87), (0)	(0.91), (0)	(0.99), (0)	(1.77), (0.11)
	LS-SCAD	13.89, 0	11.97, 0	11.28, 0	14.04, 0	13.53, 1.62
		(0.90), (0)	(0.78), (0)	(0.64), (0)	(0.99), (0)	(2.27), (0.11)
	CQR-lasso	12.15, 0	13.28, 0	12.09, 0	13.06, 0	28.26, 0.03
		(0.68), (0)	(0.62), (0)	(0.57), (0)	(0.66), (0)	(2.46), (0.02)
	CQR-SCAD	2.09, 0.02	1.29, 0	1.83, 0	1.97, 0	2.38, 0.06
		(0.40), (0.01)	(0.24), (0)	(0.32), (0)	(0.30), (0)	(0.32), (0.03)

TABLE II

SIMULATION RESULTS FOR THE NUMERICAL COMPARISON OF FOUR METHODS: LS-LASSO, LS-SCAD, CQR-LASSO AND CQR-SCAD, UNDER MODEL (15) WITH $n = 200$ AND $p = 1200$. THE LS-ORACLE AND CQR-ORACLE SERVE AS THE BENCHMARK. TWO COVARIANCE STRUCTURES $\Sigma = (0.5^{|i-j|})$ AND $\Sigma = (0.8^{|i-j|})$ ARE SHOWN, UNDER EACH OF WHICH FIVE ERROR DISTRIBUTIONS ARE CONSIDERED: $N(0,3)$, MIXTURE NORMAL, MIXTURE DOUBLE GAMMA, t_3 AND CAUCHY. THE ESTIMATION ACCURACY IS REPORTED AS "MODEL ERROR" AND THE SELECTION ACCURACY IS REPORTED AS "FP, FN". NUMBERS LISTED ARE AVERAGES OVER 100 INDEPENDENT RUNS, WITH STANDARD ERRORS REPORTED IN THE PARENTHESES

		$N(0,3)$	MN	MDG	t_3	Cauchy
			$\Sigma = (0.5^{\|i-j\|})$			
Model error	LS-oracle	0.051	0.045	0.041	0.048	1136.066
		(0.004)	(0.004)	(0.003)	(0.004)	(965.520)
	CQR-oracle	0.047	0.001	0.011	0.023	0.060
		(0.005)	(0)	(0.001)	(0.002)	(0.005)
	LS-lasso	0.340	0.337	0.281	0.284	28.450
		(0.015)	(0.014)	(0.011)	(0.013)	(7.987)
	LS-SCAD	0.061	0.061	0.055	0.062	41.685
		(0.006)	(0.005)	(0.005)	(0.005)	(24.654)
	CQR-lasso	0.394	0.072	0.180	0.239	0.830
		(0.018)	(0.011)	(0.014)	(0.013)	(0.073)
	CQR-SCAD	0.046	0.001	0.011	0.023	0.137
		(0.004)	(0)	(0.001)	(0.002)	(0.030)
FP, FN	LS-lasso	19.62, 0	20.09, 0	20.15, 0	19.59, 0	21.41, 1.66
		(1.41), (0)	(1.25), (0)	(1.22), (0)	(1.27), (0)	(3.94), (0.13)
	LS-SCAD	5.24, 0	5.76, 0	4.56, 0	6.53, 0	25.49, 1.54
		(1.08), (0)	(0.94), (0)	(0.92), (0)	(1.10), (0)	(4.01), (0.12)
	CQR-lasso	18.76, 0	19.05, 0	19.11, 0	18.59, 0	60.56, 0
		(0.95), (0)	(0.77), (0)	(0.78), (0)	(0.95), (0)	(6.35), (0)
	CQR-SCAD	2.21, 0	2.29, 0	2.99, 0	2.31, 0	1.50, 0
		(0.30), (0)	(0.48), (0)	(0.44), (0)	(0.36), (0)	(0.30), (0)
			$\Sigma = (0.8^{\|i-j\|})$			
Model error	LS-oracle	0.042	0.047	0.034	0.046	71435.826
		(0.004)	(0.005)	(0.003)	(0.004)	(68875.639)
	CQR-oracle	0.049	0.001	0.011	0.022	0.055
		(0.004)	(0)	(0.001)	(0.002)	(0.005)
	LS-lasso	0.252	0.235	0.219	0.208	22.598
		(0.012)	(0.011)	(0.009)	(0.013)	(2.334)
	LS-SCAD	0.071	0.073	0.050	0.065	23.726
		(0.006)	(0.007)	(0.004)	(0.009)	(2.856)
	CQR-lasso	0.255	0.030	0.099	0.153	0.730
		(0.014)	(0.004)	(0.008)	(0.009)	(0.070)
	CQR-SCAD	0.086	0.001	0.023	0.048	0.539
		(0.008)	(0)	(0.003)	(0.004)	(0.099)
FP, FN	LS-lasso	14.83, 0	16.17, 0	15.24, 0	14.80, 0	20.44, 1.63
		(1.03), (0)	(1.1), (0)	(1.09), (0)	(1.23), (0)	(3.83), (0.12)
	LS-SCAD	6.36, 0	5.68, 0	5.64, 0	4.74, 0.01	15.22, 1.84
		(0.99), (0)	(0.72), (0)	(0.83), (0)	(0.76), (0.01)	(2.92), (0.10)
	CQR-lasso	16.86, 0	14.89, 0	15.83, 0	16.47, 0	42.75, 0
		(1.02), (0)	(0.7), (0)	(0.78), (0)	(0.98), (0)	(4.13), (0)
	CQR-SCAD	2.19, 0	1.93, 0	2.70, 0	2.59, 0	1.85, 0
		(0.32), (0)	(0.36), (0)	(0.50), (0)	(0.44), (0)	(0.24), (0)

TABLE III

SIMULATION RESULTS FOR NUMERICAL COMPARISON BETWEEN CV, BIC^H AND BIC^{HL} IN TERMS OF TUNING PARAMETER SELECTION, UNDER MODEL (15) WITH $n = 100$ AND $p = 600$. TWO COVARIANCE STRUCTURES $\Sigma = (0.5^{|i-j|})$ AND $\Sigma = (0.8^{|i-j|})$ ARE SHOWN, UNDER EACH OF WHICH FIVE ERROR DISTRIBUTIONS ARE CONSIDERED: $N(0,3)$, MIXTURE NORMAL, MIXTURE DOUBLE GAMMA, t_3 AND CAUCHY. THE ESTIMATION ACCURACY IS REPORTED AS "MODEL ERROR" AND THE SELECTION ACCURACY IS REPORTED AS "FP, FN". NUMBERS LISTED ARE AVERAGES OVER 100 INDEPENDENT RUNS, WITH STANDARD ERRORS REPORTED IN THE PARENTHESES

		$N(0,3)$	MN	MDG	t_3	Cauchy
			$\Sigma = (0.5^{\|i-j\|})$			
Model error	CV	0.087	0.005	0.021	0.043	0.312
		(0.008)	(0.002)	(0.002)	(0.004)	(0.083)
	BIC^{H}	0.115	0.017	0.053	0.058	0.988
		(0.019)	(0.011)	(0.017)	(0.012)	(0.374)
	BIC^{HL}	0.094	0.003	0.022	0.048	6.287
		(0.011)	(0.001)	(0.003)	(0.004)	(0.824)
FP, FN	CV	0, 0	0, 0	0, 0	0, 0	0.01, 0.13
		(0), (0)	(0), (0)	(0), (0)	(0), (0)	(0.01), (0.04)
	BIC^{H}	0, 0.03	0, 0.01	0, 0.03	0, 0.01	0.01, 0.27
		(0), (0.02)	(0), (0.01)	(0), (0.02)	(0), (0.01)	(0.01), (0.05)
	BIC^{HL}	0, 0.01	0, 0	0, 0	0, 0	0, 1.17
		(0), (0.01)	(0), (0)	(0), (0)	(0), (0)	(0), (0.09)
			$\Sigma = (0.8^{\|i-j\|})$			
Model error	CV	0.273	0.008	0.088	0.118	0.718
		(0.047)	(0.003)	(0.025)	(0.025)	(0.116)
	BIC^{H}	0.393	0.077	0.200	0.295	2.169
		(0.066)	(0.035)	(0.051)	(0.059)	(0.338)
	BIC^{HL}	0.607	0.024	0.153	0.114	5.411
		(0.091)	(0.020)	(0.043)	(0.020)	(0.517)
FP, FN	CV	0, 0.07	0, 0	0.02, 0.02	0 , 0.01	0.1, 0.2
		(0), (0.03)	(0), (0)	(0.02), (0.01)	(0), (0.01)	(0.03), (0.04)
	BIC^{H}	0.01, 0.15	0, 0.04	0, 0.09	0, 0.12	0.14, 0.53
		(0.01), (0.04)	(0), (0.02)	(0), (0.03)	(0), (0.03)	(0.06), (0.07)
	BIC^{HL}	0.05, 0.23	0 , 0.01	0, 0.06	0, 0.01	0.12, 1.07
		(0.02), (0.04)	(0), (0.01)	(0), (0.02)	(0), (0.01)	(0.04), (0.08)

TABLE IV

SIMULATION RESULTS FOR NUMERICAL COMPARISON BETWEEN CV, BIC^H AND BIC^{HL} IN TERMS OF TUNING PARAMETER SELECTION, UNDER MODEL (15) WITH $n = 200$ AND $p = 1200$. TWO COVARIANCE STRUCTURES $\Sigma = (0.5^{|i-j|})$ AND $\Sigma = (0.8^{|i-j|})$ ARE SHOWN, UNDER EACH OF WHICH FIVE ERROR DISTRIBUTIONS ARE CONSIDERED: $N(0,3)$, MIXTURE NORMAL, MIXTURE DOUBLE GAMMA, t_3 AND CAUCHY. THE ESTIMATION ACCURACY IS REPORTED AS "MODEL ERROR" AND THE SELECTION ACCURACY IS REPORTED AS "FP, FN". NUMBERS LISTED ARE AVERAGES OVER 100 INDEPENDENT RUNS, WITH STANDARD ERRORS REPORTED IN THE PARENTHESES

		$N(0,3)$	MN	MDG	t_3	Cauchy
			$\Sigma = (0.5^{\|i-j\|})$			
Model error	CV	0.121	0.006	0.026	0.043	0.477
		(0.019)	(0.002)	(0.003)	(0.006)	(0.224)
	BIC^{H}	0.041	0.001	0.012	0.020	0.915
		(0.004)	(0.000)	(0.001)	(0.002)	(0.423)
	BIC^{HL}	0.042	0.001	0.009	0.023	3.632
		(0.004)	(0.000)	(0.001)	(0.002)	(0.757)
FP, FN	CV	0, 0.02	0, 0	0, 0	0, 0	0.01, 0.11
		(0), (0.01)	(0), (0)	(0), (0)	(0), (0)	(0.01), (0.04)
	BIC^{H}	0, 0	0, 0	0, 0	0, 0	0.26, 0.05
		(0), (0)	(0), (0)	(0), (0)	(0), (0)	(0.13), (0.03)
	BIC^{HL}	0, 0	0, 0	0, 0	0, 0	0.01, 0.61
		(0), (0)	(0), (0)	(0), (0)	(0), (0)	(0.01), (0.08)
			$\Sigma = (0.8^{\|i-j\|})$			
Model error	CV	0.372	0.042	0.081	0.206	0.736
		(0.059)	(0.025)	(0.023)	(0.044)	(0.141)
	BIC^{H}	0.235	0.001	0.070	0.055	0.538
		(0.05)	(0.000)	(0.03)	(0.017)	(0.086)
	BIC^{HL}	0.132	0.002	0.019	0.041	3.28
		(0.032)	(0.001)	(0.002)	(0.004)	(0.518)
FP, FN	CV	0.01, 0.12	0, 0.02	0.01, 0.02	0, 0.06	0.05, 0.16
		(0.01), (0.03)	(0), (0.01)	(0.01), (0.01)	(0), (0.02)	(0.02), (0.04)
	BIC^{H}	0, 0.09	0, 0	0, 0.03	0, 0.01	0.21, 0.12
		(0), (0.03)	(0), (0)	(0), (0.02)	(0), (0.01)	(0.06), (0.03)
	BIC^{HL}	0, 0.03	0, 0	0, 0	0, 0	0.10, 0.55
		(0), (0.02)	(0), (0)	(0), (0)	(0), (0)	(0.04), (0.07)

VIII. PROOFS

We provide proofs of all previously stated results in this section. For the sake of brevity, some auxiliary results are relegated to the appendix.

Assume without loss of generality $\beta_0^* = 0$ such that $\alpha_k^* = F^{-1}(\tau_k)$ for $1 \le k \le K$. Also recall $M_0 = \max_{0\le j\le p} \|X_j\|_2^2/n$ from Theorem 1.

Lemma 3. *Under condition (C0), with probability at least*

$$1 - 2K\exp\Big(-\frac{9}{2}n\lambda^2\Big) - 2p\exp\Big(-\frac{n\lambda^2}{2M_0}\Big),$$

the CQR lasso estimator $(\widehat{\alpha}_\lambda, \widehat{\beta}_\lambda)$ *satisfies*

$$(\widehat{\delta}^\lambda, \widehat{\Delta}^\lambda) \in \mathscr{C} = \{(\delta,\Delta) \in \mathbb{R}^K \times \mathbb{R}^p \colon \|\Delta_{\mathscr{A}^c}\|_1 \le 3\|\Delta_{\mathscr{A}}\|_1 + \tfrac{3}{K}\|\delta\|_1\},$$

where $\widehat{\delta}^\lambda = \widehat{\alpha}_\lambda - \alpha^*$ *and* $\widehat{\Delta}^\lambda = \widehat{\beta}_\lambda - \beta^*$.

Proof of Lemma 3. See Section A of the appendix. □

Now let $v_n(\alpha,\beta) = Q_n(\alpha,\beta) - Q_n(\alpha^*,\beta^*) - \mathbb{E}\big[Q_n(\alpha,\beta) - Q_n(\alpha^*,\beta^*)\big]$. For $r > 0$, set $\mathscr{C}_r = \{(\delta,\Delta) \in \mathscr{C} \colon (nK)^{-1} \sum_{k=1}^K \sum_{i=1}^n (\delta_k + \mathbf{x}_i^{\mathrm{T}}\Delta)^2 \le r^2\}$ and define $e(r) = \sup_{(\delta,\Delta)\in\mathscr{C}_r} |v_n(\alpha^* + \delta, \beta^* + \Delta)|$.

Lemma 4. *For* $r, t > 0$, *under conditions (C0) and (C1), with probability at least* $1 - \exp\big[-nt^2/(32r^2)\big]$, *we have*

$$e(r) \le 16\sqrt{\frac{2M_0}{\kappa_0}}\sqrt{\frac{1+\log p}{n}}\big(\sqrt{s}+1\big)r + t.$$

It follows immediately that, if one takes

$$t = 16\sqrt{\frac{2M_0}{\kappa_0}}\sqrt{\frac{1+\log p}{n}}\big(\sqrt{s}+1\big)r,$$

then with probability at least $1 - \exp\big[-2M_0\kappa_0^{-1}s(1+\log p)\big]$, *we have*

$$e(r) \le 32\sqrt{\frac{2M_0}{\kappa_0}}\sqrt{\frac{1+\log p}{n}}\big(\sqrt{s}+1\big)r.$$

Proof of Lemma 4. See Section A of the appendix. □

Lemma 5. *Under conditions (C0) and (C2), for any* $(\delta,\Delta) \in \mathscr{C}$, *we have*

$$\mathbb{E}\big[Q_n(\alpha^* + \delta, \beta^* + \Delta) - Q_n(\alpha^*,\beta^*)\big] \ge \min\big\{\underline{f}r^2/4,\, q\big(\underline{f}/K\big)^{1/2}r\big\},$$

where $r^2 = (nK)^{-1}\sum_{i=1}^n \sum_{k=1}^K (\delta_k + \mathbf{x}_i^T\Delta)^2$.

Proof of Lemma 5. See Section A of the appendix. □

Proof of Theorem 1. Let

$$r_* = 8\underline{f}^{-1}\left[32\sqrt{\frac{2M_0}{\kappa_0}}\sqrt{\frac{1+\log p}{n}}\,(\sqrt{s}+1) + \lambda\sqrt{\frac{s}{\kappa_0}}\right]$$

and set $\mathscr{C}^* = \{(\delta,\Delta)\in\mathscr{C}\colon (nK)^{-1}\sum_{i=1}^n\sum_{k=1}^K(\delta_k+\mathbf{x}_i^{\mathrm{T}}\Delta)^2 = r_*^2\}$. Moreover, define $\widehat{\delta}^{\lambda} = \widehat{\alpha}_\lambda - \alpha^*$ and $\widehat{\Delta}^{\lambda} = \widehat{\beta}_\lambda - \beta^*$. Under event $\mathscr{E}_1 = \{(\widehat{\delta}^{\lambda},\widehat{\Delta}^{\lambda})\in\mathscr{C}\}$, if

$$\inf_{(\delta,\Delta)\in\mathscr{C}^*} Q_n(\alpha^*+\delta,\beta^*+\Delta) - Q_n(\alpha^*,\beta^*) + \lambda(\|\beta^*+\Delta\|_1 - \|\beta^*\|_1) > 0, \tag{16}$$

then by convexity of Q_n, this implies that $(\widehat{\delta}^{\lambda},\widehat{\Delta}^{\lambda})\in\mathscr{C}_{r_*}$. To show (16), first note that for all $(\delta,\Delta)\in\mathscr{C}^*$,

$$\begin{aligned} &Q_n(\alpha^*+\delta,\beta^*+\Delta) - Q_n(\alpha^*,\beta^*) + \lambda(\|\beta^*+\Delta\|_1 - \|\beta^*\|_1) \\ &\quad\ge \mathbb{E}\big[Q_n(\alpha^*+\delta,\beta^*+\Delta) - Q_n(\alpha^*,\beta^*)\big] - e(r_*) \\ &\qquad + \lambda\big(\|\Delta_{\mathscr{A}^{\mathrm{c}}}\|_1 - \|\Delta_{\mathscr{A}}\|_1\big). \end{aligned} \tag{17}$$

Now let $\mathscr{E}_2 = \big\{e(r_*)\le 32\sqrt{2M_0(1+\log p)/(n\kappa_0)}\,(\sqrt{s}+1)r_*\big\}$. It follows from Lemma 4 that $\Pr(\mathscr{E}_2)\ge 1-\exp\big[-2M_0\kappa_0^{-1}s(1+\log p)\big]$. By Lemma 5, for any $(\delta,\Delta)\in\mathscr{C}^*$, we have

$$\mathbb{E}\big[Q_n(\alpha^*+\delta,\beta^*+\Delta) - Q_n(\alpha^*,\beta^*)\big] \ge \min\big\{\underline{f}r_*^2/4,\, q(\underline{f}/K)^{1/2}r_*\big\}.$$

Also, by condition (C1) and (A.5), for $(\delta,\Delta)\in\mathscr{C}^*$, we have $\|\Delta_{\mathscr{A}}\|_1\le r_*\sqrt{s/\kappa_0}$. Thus, under event $\mathscr{E}_1\cap\mathscr{E}_2$, for any $(\delta,\Delta)\in\mathscr{C}^*$, it follows from (17) and the growth condition that

$$\begin{aligned} &Q_n(\alpha^*+\delta,\beta^*+\Delta) - Q_n(\alpha^*,\beta^*) + \lambda(\|\beta^*+\Delta\|_1 - \|\beta^*\|_1) \\ &\ge \frac{\underline{f}}{4}r_*^2 - \left[32\sqrt{\frac{2M_0}{\kappa_0}}\sqrt{\frac{1+\log p}{n}}\,(\sqrt{s}+1) + \lambda\sqrt{\frac{s}{\kappa_0}}\right]r_* > 0 \end{aligned}$$

by our choice of r_*. Therefore, by Lemma 3 and 4, with probability at least

$$\Pr(\mathscr{E}_1\cap\mathscr{E}_2)\ge 1-\Pr(\mathscr{E}_1^{\mathrm{c}})-\Pr(\mathscr{E}_2^{\mathrm{c}})\ge 1-p_1(\lambda),$$

we have $(\widehat{\delta}^{\lambda},\widehat{\Delta}^{\lambda})\in\mathscr{C}_{r_*}$. This, by condition (C1), further implies that

$$\begin{aligned} r_*^2 &\ge \frac{\kappa_m}{K}\Big[\|\hat{\delta}^{\lambda}\|_2^2 + K\|\widehat{\Delta}^{\lambda}_{\mathscr{A}\cup\overline{\mathscr{A}}(\widehat{\Delta}^{\lambda},m)}\|_2^2\Big] \\ &\ge \frac{\kappa_m}{K}\|\widehat{\delta}^{\lambda}\|_2^2 + \kappa_m\|\widehat{\Delta}^{\lambda}_{\mathscr{A}\cup\overline{\mathscr{A}}(\widehat{\Delta}^{\lambda},m)}\|_2^2. \end{aligned}$$

As a result, we obtain $\|\widehat{\delta}^{\lambda}\|_2\le r_*\sqrt{K/\kappa_m}$ and

$$\|\widehat{\Delta}^{\lambda}_{\mathscr{A}\cup\overline{\mathscr{A}}(\widehat{\Delta}^{\lambda},m)}\|_2 \le r_*/\sqrt{\kappa_m}. \tag{18}$$

Note that the jth largest in absolute value component of $\widehat{\Delta}_{\mathscr{A}^{\mathrm{c}}}$ is bounded by $\|\widehat{\Delta}_{\mathscr{A}^{\mathrm{c}}}\|_1/j$. Therefore, it follows that

$$\begin{aligned} &\Big\|\widehat{\Delta}_{(\mathscr{A}\cup\overline{\mathscr{A}}(\widehat{\Delta}^{\lambda},m))^{\mathrm{c}}}\Big\|_2^2 \le \sum_{j=m+1}^{p}\frac{\|\widehat{\Delta}^{\lambda}_{\mathscr{A}^{\mathrm{c}}}\|_1^2}{j^2} \le \frac{1}{m}\|\widehat{\Delta}^{\lambda}_{\mathscr{A}^{\mathrm{c}}}\|_1^2 \\ &\le \frac{1}{m}\big[3\|\widehat{\Delta}^{\lambda}_{\mathscr{A}}\|_1 + 3K^{-1}\|\widehat{\delta}^{\lambda}\|_1\big]^2 \le \frac{18s}{m}\|\widehat{\Delta}^{\lambda}_{\mathscr{A}}\|_2^2 + \frac{18}{mK}\|\widehat{\delta}^{\lambda}\|_2^2 \\ &\le \frac{18s}{m}\|\widehat{\Delta}^{\lambda}_{\mathscr{A}\cup\overline{\mathscr{A}}(\widehat{\Delta}^{\lambda},m)}\|_2^2 + \frac{18}{mK}\|\widehat{\delta}^{\lambda}\|_2^2, \end{aligned}$$

which, together with (18), implies that

$$\begin{aligned} \|\widehat{\Delta}^{\lambda}\|_2^2 &\le \Big(1+\frac{18s}{m}\Big)\|\widehat{\Delta}^{\lambda}_{\mathscr{A}\cup\overline{\mathscr{A}}(\widehat{\Delta}^{\lambda},m)}\|_2^2 + \frac{18}{mK}\|\widehat{\delta}^{\lambda}\|_2^2 \\ &\le \frac{r_*^2}{\kappa_m}\Big(1+\frac{18s}{m}+\frac{18}{m}\Big). \end{aligned}$$

This completes the proof of Theorem 1. □

For $r>0$, define $B_{\mathscr{A}}(r)=\{(\delta,\Delta)\in\mathbb{R}^K\times\mathbb{R}^p\colon \|\delta\|_2^2+\|\Delta_{\mathscr{A}}\|_2^2\le r^2, \Delta_{\mathscr{A}^c}=\mathbf{0}\}$ and $S_{\mathscr{A}}(r)=\{(\delta,\Delta)\in\mathbb{R}^K\times\mathbb{R}^p\colon \|\delta\|_2^2+\|\Delta_{\mathscr{A}}\|_2^2=r^2, \Delta_{\mathscr{A}^c}=\mathbf{0}\}$. Moreover, let $z(r)=\sup_{(\delta,\Delta)\in B_{\mathscr{A}}(r)}|\nu_n(\alpha^*+\delta,\beta^*+\Delta)|$.

Lemma 6. *Under condition (C0), for any $r,t>0$ such that $r\sqrt{(s+1)M_{\mathscr{A}}}\le\mathscr{U}_0$, with probability at least $1-\exp\big[-nt^2/(32\bar{\mu}r^2)\big]$, we have*

$$z(r)\le 4r\sqrt{\frac{(s+1)M_{\mathscr{A}}}{n}}+t$$

and

$$\inf_{(\delta,\Delta)\in S_{\mathscr{A}}(r)}\big[Q_n(\alpha^*+\delta,\beta^*+\Delta)-Q_n(\alpha^*,\beta^*)\big]\ge\frac{\underline{f}}{2K}\underline{\mu}r^2-4r\sqrt{\frac{(s+1)M_{\mathscr{A}}}{n}}-t.$$

Proof of Lemma 6. See Section A of the appendix. □

Proof of Lemma 1. For ease of notation, let $\widehat{\delta}^{\mathrm{o}}=\widehat{\alpha}^{\mathrm{o}}-\alpha^*$ and $\widehat{\Delta}^{\mathrm{o}}=\widehat{\beta}^{\mathrm{o}}-\beta^*$. In Lemma 6, let $r=r^*=32K(\underline{f}\,\underline{\mu})^{-1}\times\sqrt{M_{\mathscr{A}}(s+1)/n}$ and $t=4r^*\sqrt{M_{\mathscr{A}}(s+1)/n}$. By assumption, with the choice of r^*, we have $r^*\sqrt{(s+1)M_{\mathscr{A}}}\le\mathscr{U}_0$. It follows immediately that with probability at least $1-\exp\big[-(s+1)M_{\mathscr{A}}/(2\bar{\mu})\big]$, we have

$$\inf_{(\delta,\Delta)\in S_{\mathscr{A}}(r^*)}\big[Q_n(\alpha^*+\delta,\beta^*+\Delta)-Q_n(\alpha^*,\beta^*)\big]\ge\frac{\underline{f}\,\underline{\mu}(r^*)^2}{2K}-8r^*\sqrt{\frac{(s+1)M_{\mathscr{A}}}{n}}>0.$$

By convexity of Q_n and optimality of $(\widehat{\alpha}^{\mathrm{o}},\widehat{\beta}^{\mathrm{o}})$, this implies that

$$\|\widehat{\delta}^{\mathrm{o}}\|_2^2+\|\widehat{\Delta}^{\mathrm{o}}\|_2^2\le(r^*)^2,$$

which completes the lemma. □

Lemma 7. *Suppose the folded concave penalized CQR* (6) *is solved with the LLA algorithm. Let $a_0=\min(a_2,1)$ and define*

$$\mathscr{E}_1=\{\|\widehat{\beta}^{(0)}-\beta^*\|_\infty\le a_0\lambda\},$$
$$\mathscr{E}_2=\{\|\nabla_{\mathscr{A}^c}Q_n(\widehat{\alpha}^o,\widehat{\beta}^o)\|_\infty<a_1\lambda\},$$
$$\mathscr{E}_3=\Big\{\min_{j\in\mathscr{A}}|\hat{\beta}_j^o|>a\lambda\Big\},$$

where $\nabla_{\mathscr{A}^c}Q_n(\widehat{\alpha}^o,\widehat{\beta}^o)=\big(\nabla_jQ_n(\widehat{\alpha}^o,\widehat{\beta}^o),\,j\in\mathscr{A}^c\big)$ with

$$\nabla_jQ_n(\widehat{\alpha}^o,\widehat{\beta}^o)=\frac{1}{2n}\sum_{i=1}^n x_{ij}\Big(1-\frac{2}{K}\sum_{k=1}^K\tau_k\Big)-\frac{1}{2nK}\sum_{i=1}^n\sum_{k=1}^K Sgn(\hat{r}_{ik})x_{ij},$$

$\hat{r}_{ik}=y_i-\hat{\alpha}_k^o-\mathbf{x}_i^T\widehat{\beta}^o$, $1\le i\le n$, $1\le k\le K$, *and*

$$Sgn(u)=\begin{cases}1, & \text{if } u>0\\ [-1,1], & \text{if } u=0\\ -1, & \text{if } u<0.\end{cases}$$

Then under $\mathscr{E}_1\cap\mathscr{E}_2\cap\mathscr{E}_3$ and condition (C0), the LLA algorithm converges to the CQR oracle estimator.

Proof of Lemma 7. See Section A of the appendix. □

For each $j \in \mathscr{A}^{\mathrm{c}}$, define $S_j^n(\alpha,\beta) = \frac{1}{nK}\sum_{i=1}^n\sum_{k=1}^K \big[I(y_i-\alpha_k-\mathbf{x}_i^{\mathrm{T}}\beta\le 0)-\tau_k\big]x_{ij}$, and for $r>0$, let

$$\gamma_j(r) = \sup_{(\delta,\Delta)\in B_{\mathscr{A}}(r)} \big|S_j^n(\alpha^*+\delta,\beta^*+\Delta) - S_j^n(\alpha^*,\beta^*) - \mathbb{E}\big[S_j^n(\alpha^*+\delta,\beta^*+\Delta) - S_j^n(\alpha^*,\beta^*)\big]\big|.$$

Lemma 8. *For $r,t>0$, $0<\psi<r$ and $j\in\mathscr{A}^c$, under condition (C0) we have*

$$\Pr(\gamma_j(r)>t) \le 2N_\psi \exp\bigg(-\frac{nt^2}{8\bar{f}M_{\mathscr{A}^c}^2\bar{\mu}^{1/2}r+\frac{8}{3}M_{\mathscr{A}^c}t}\bigg) + 2N_\psi\exp\bigg(-\frac{nt_0^2}{2\bar{f}M_{\mathscr{A}^c}^2\big((s+1)M_{\mathscr{A}}\big)^{1/2}\psi+\frac{4}{3}M_{\mathscr{A}^c}t_0}\bigg),$$

where N_ψ is the ψ-covering number (see, e.g., [24]) of $B_{\mathscr{A}}(r)$ and $t_0 = \big[t/2-2\bar{f}M_{\mathscr{A}^c}\big((s+1)M_{\mathscr{A}}\big)^{1/2}\psi\big]_+$.

Proof of Lemma 8. See Section A of the appendix. □

Proof of Theorem 2. Let $\widehat{\delta}^{\mathrm{o}} = \widehat{\alpha}^{\mathrm{o}}-\alpha^*$ and $\widehat{\Delta}^{\mathrm{o}} = \widehat{\beta}^{\mathrm{o}}-\beta^*$. For $1\le i\le n$, $1\le k\le K$, write $\hat{r}_{ik} = y_i-\hat{\alpha}_k^{\mathrm{o}}-\mathbf{x}_i^{\mathrm{T}}\widehat{\beta}^{\mathrm{o}}$ and $r_{ik}^* = y_i-\alpha_k^*-\mathbf{x}_i^{\mathrm{T}}\beta^*$. For ease of notation, let $F(\delta,\Delta) = Q_n(\alpha^*+\delta,\beta^*+\Delta)-Q_n(\alpha^*,\beta^*)$ for $(\delta,\Delta)\in\mathbb{R}^K\times\mathbb{R}^p$. According to Lemma 7, with probability at least

$$\Pr(\mathscr{E}_1\cap\mathscr{E}_2\cap\mathscr{E}_3)\ge 1-\Pr(\mathscr{E}_1^{\mathrm{c}})-\Pr(\mathscr{E}_2^{\mathrm{c}})-\Pr(\mathscr{E}_3^{\mathrm{c}}),$$

the LLA algorithm will converge to the oracle estimator in two iterations. In the sequel, we will split the proof into three parts and provide the upper bound on each of $\Pr(\mathscr{E}_1^{\mathrm{c}})$, $\Pr(\mathscr{E}_2^{\mathrm{c}})$ and $\Pr(\mathscr{E}_3^{\mathrm{c}})$, separately.

(i) First, we deal with $\Pr(\mathscr{E}_1^{\mathrm{c}}) = \Pr(\|\widehat{\beta}^{(0)}-\beta^*\|_\infty > a_0\lambda)$. Since in the LLA algorithm, we take $(\widehat{\alpha}^{(0)},\widehat{\beta}^{(0)})$ to be the lasso estimator $(\widehat{\alpha}_{\lambda_0},\widehat{\beta}_{\lambda_0})$, by Theorem 1, we have

$$\Pr(\mathscr{E}_1) = \Pr\big(\|\widehat{\beta}_{\lambda_0}-\beta^*\|_\infty\le a_0\lambda\big)\ge\Pr\big(\|\widehat{\beta}_{\lambda_0}-\beta^*\|_2\le a_0\lambda\big)\ge 1-p_1(\lambda_0),$$

which implies that $\Pr(\mathscr{E}_1^{\mathrm{c}})\le p_1(\lambda_0)$.

(ii) We next derive the upper bound on $\Pr(\mathscr{E}_3^{\mathrm{c}}) = \Pr\big(\min_{j\in\mathscr{A}}|\hat{\beta}_j^{\mathrm{o}}|\le a\lambda\big)$. Let $r_0 = \min_{j\in\mathscr{A}}|\beta_j^*|-a\lambda$. It can be seen that $\Pr(\mathscr{E}_3^{\mathrm{c}})\le\Pr(\|\widehat{\Delta}^{\mathrm{o}}\|_\infty>r_0)$. Note that by convexity of Q_n, $\|\widehat{\Delta}^{\mathrm{o}}\|_2\le r_0$ is implied by the event that $\inf_{(\delta,\Delta)\in S_{\mathscr{A}}(r_0)}F(\delta,\Delta)>0$. Since $r_0\sqrt{(s+1)M_{\mathscr{A}}}\le\mathscr{U}_0$, it follows from Lemma 6 that for any $t>0$,

$$\inf_{(\delta,\Delta)\in S_{\mathscr{A}}(r_0)}F(\delta,\Delta)\ge\frac{\underline{f}}{2K}\underline{\mu}r_0^2-4r_0\sqrt{\frac{(s+1)M_{\mathscr{A}}}{n}}-t$$

holds with probability at least $1-\exp\big[-nt^2/(32\bar{\mu}r_0^2)\big]$. By condition (C3), it can be seen that $r_0>\lambda>8K(\underline{f}\,\underline{\mu})^{-1}\times\sqrt{(s+1)M_{\mathscr{A}}/n}$. Now take $t=\underline{f}\,\underline{\mu}r_0^2/(4K)-2r_0\sqrt{(s+1)M_{\mathscr{A}}/n}$. Then, we can see that $t>0$. It follows immediately that $\inf_{(\delta,\Delta)\in S_{\mathscr{A}}(r_0)}F(\delta,\Delta)\ge t>0$. With this specific choice of t, we get

$$\Pr\big(\|\widehat{\Delta}^{\mathrm{o}}\|_2\le r_0\big)\ge 1-\exp\big[-nt^2/(32\bar{\mu}r_0^2)\big],$$

which implies that

$$\Pr(\mathscr{E}_3^{\mathrm{c}})\le\Pr(\|\widehat{\Delta}^{\mathrm{o}}\|_\infty>r_0)\le\Pr(\|\widehat{\Delta}^{\mathrm{o}}\|_2>r_0)\le\exp\big[-nt^2/(32\bar{\mu}r_0^2)\big].$$

(iii) Finally, we look at $\Pr(\mathscr{E}_2^{\rm c}) = \Pr(\|\nabla_{\mathscr{A}^{\rm c}} Q_n(\widehat{\alpha}^{\rm o}, \widehat{\beta}^{\rm o})\|_\infty \geq a_1\lambda)$. To this end, we set $r_* = \sqrt{(s+1)M_{\mathscr{A}} \log n/n}$ and let $\mathscr{R} = \{(i,k)\colon \hat{r}_{ik} = 0,\, 1 \leq i \leq n,\, 1 \leq k \leq K\}$ be the index set of zero residuals. From Section B of the appendix, we have $|\mathscr{R}| \leq K(K+s)$. It follows that

$$\begin{aligned}\nabla_j Q_n(\widehat{\alpha}^{\rm o}, \widehat{\beta}^{\rm o}) &= \frac{1}{2nK} \sum_{i=1}^n \sum_{k=1}^K \big[(1-2\tau_k) - \mathrm{Sgn}(\hat{r}_{ik})\big] x_{ij} \\ &= \frac{1}{nK} \sum_{i=1}^n \sum_{k=1}^K \big[I(\hat{r}_{ik} \leq 0) - \tau_k\big] x_{ij} - \frac{1}{2nK} \sum_{(i,k)\in\mathscr{R}} \big[\mathrm{Sgn}(\hat{r}_{ik}) + 1\big] x_{ij},\end{aligned}$$

where we have

$$\max_{j\in\mathscr{A}^{\rm c}} \bigg| \frac{1}{2nK} \sum_{(i,k)\in\mathscr{R}} \big[\mathrm{Sgn}(\hat{r}_{ik}) + 1\big] x_{ij} \bigg| \leq \frac{(K+s)M_{\mathscr{A}^{\rm c}}}{n} := B_1.$$

Now define event $\mathscr{E}_0 = \{(\widehat{\delta}^{\rm o}, \widehat{\Delta}^{\rm o}) \in B_{\mathscr{A}}(r_*)\}$. Under $\mathscr{E}_0$, by the triangular inequality, we have

$$\begin{aligned}\max_{j\in\mathscr{A}^{\rm c}} \bigg| \frac{1}{nK} \sum_{i=1}^n \sum_{k=1}^K \big[I(\hat{r}_{ik} \leq 0) - \tau_k\big] x_{ij} \bigg| &\leq \max_{j\in\mathscr{A}^{\rm c}} \gamma_j(r_*) + \max_{j\in\mathscr{A}^{\rm c}} |S_j^n(\alpha^*, \beta^*)| \\ &+ \max_{j\in\mathscr{A}^{\rm c}} \sup_{(\delta,\Delta)\in B_{\mathscr{A}}(r_*)} \Big| \mathbb{E}\big[S_j^n(\alpha^*+\delta, \beta^*+\Delta) - S_j^n(\alpha^*, \beta^*)\big] \Big|.\end{aligned}$$

By the mean value theorem, it can be seen that

$$\begin{aligned}&\max_{j\in\mathscr{A}^{\rm c}} \sup_{(\delta,\Delta)\in B_{\mathscr{A}}(r_*)} \Big| \mathbb{E}\big[S_j^n(\alpha^*+\delta, \beta^*+\Delta) - S_j^n(\alpha^*, \beta^*)\Big| \\ &\leq \frac{1}{nK} \bar{f} M_{\mathscr{A}^{\rm c}} \sup_{(\delta,\Delta)\in B_{\mathscr{A}}(r_*)} \sum_{i=1}^n \sum_{k=1}^K |\delta_k + \mathbf{x}_{i\mathscr{A}}^{\rm T} \Delta_{\mathscr{A}}| \leq \bar{f} M_{\mathscr{A}^{\rm c}} \bar{\mu}^{1/2} r_* := B_2.\end{aligned}$$

Note that if $B \geq 0$, then $2B = a_1\lambda - B_1 - B_2$. It follows that

$$\begin{aligned}\Pr(\mathscr{E}_2^{\rm c}) \leq \Pr\big((\widehat{\delta}^{\rm o}, \widehat{\Delta}^{\rm o}) \notin B_{\mathscr{A}}(r_*)\big) &+ \Pr\Big(\max_{i\in\mathscr{A}^{\rm c}} \gamma_j(r_*) \geq B\Big) \\ &+ \Pr\Big(\max_{j\in\mathscr{A}^{\rm c}} |S_j^n(\alpha^*, \beta^*)| \geq B\Big).\end{aligned}$$

Note that $r_* \sqrt{(s+1)M_{\mathscr{A}}} \leq \mathscr{U}_0$. By similar arguments in (ii), it can be shown that

$$\Pr\big((\widehat{\delta}^{\rm o}, \widehat{\Delta}^{\rm o}) \notin B_{\mathscr{A}}(r_*)\big) \leq \exp\bigg(-\frac{nt_*^2}{32\bar{\mu} r_*^2}\bigg),$$

where $t_* = \underline{f}\, \underline{\mu} r_*^2/(4K) - 2r_* \sqrt{(s+1)M_{\mathscr{A}}/n}$. Applying Hoeffding's inequality, we obtain

$$\Pr\Big(\max_{j\in\mathscr{A}^{\rm c}} |S_j^n(\alpha^*, \beta^*)| \geq B\Big) \leq 2(p-s) \exp\bigg(-\frac{2nB^2}{M_0}\bigg).$$

Lastly, we apply Lemma 8 to obtain the bound on $\Pr\Big(\max_{j\in\mathscr{A}^{\rm c}} \gamma_j(r_*) \geq B\Big)$. Let $\psi = 4r_*/n^2$. It can be shown that the ψ-covering number of $B_{\mathscr{A}}(r_*)$ satisfies

$$\Big(\frac{r_*}{\psi}\Big)^{K+s} \leq N_\psi \leq \Big(\frac{2r_* + \psi}{\psi}\Big)^{K+s} \leq n^{2(K+s)},\, n \geq 2.$$

By Lemma 8, we have

$$\begin{aligned}\Pr\Big(\max_{j\in\mathscr{A}^{\rm c}} \gamma_j(r_*) \geq B\Big) &\leq 2(p-s) N_\psi \exp\bigg(-\frac{nB^2}{8\bar{f} M_{\mathscr{A}^{\rm c}}^2 \bar{\mu}^{1/2} r_* + \frac{8}{3} M_{\mathscr{A}^{\rm c}} B}\bigg) \\ &+ 2(p-s) N_\psi \exp\bigg(-\frac{nB_0^2}{2\bar{f} M_{\mathscr{A}^{\rm c}}^2 \big((s+1)M_{\mathscr{A}}\big)^{1/2} \psi + \frac{4}{3} M_{\mathscr{A}^{\rm c}} B_0}\bigg),\end{aligned}$$

where $B_0 = \big[B/2 - 2\bar{f} M_{\mathscr{A}^c}\big((s+1)M_{\mathscr{A}}\big)^{1/2}\psi\big]_+$. This completes the proof. □

Lemma 9. *For any $A \in \{S : S \supset \mathscr{A}, |S| \le 2J_n\}$ and $r,t > 0$, let $z(A,r) = \sup_{(\delta,\Delta)\in B_A(r)} |\nu_n(\alpha^* + \delta, \beta^* + \Delta)|$. With probability at least $1 - \exp\{-nt^2/(32\bar{\zeta} r^2)\}$, we have*

$$z(A,r) \le 4Mr\sqrt{\frac{|A|+1}{n}} + t.$$

Proof of Lemma 9. See Section A of the appendix. □

Proof of Theorem 3. Split all models (denoted by their index sets) under consideration, $\{\hat{A}_\lambda : \lambda \in \Xi_n\}$, into three groups: $\{\hat{A}_\lambda : \lambda \in \Xi_n^-\}$, $\{\hat{A}_\lambda : \lambda \in \Xi_n^0\}$, and $\{\hat{A}_\lambda : \lambda \in \Xi_n^+\}$, where $\Xi_n^- = \{\lambda \in \Xi_n : \mathscr{A} \not\subset \hat{A}_\lambda\}$ (underfitted models), $\Xi_n^0 = \{\lambda \in \Xi_n : \hat{A}_\lambda = \mathscr{A}\}$ and $\Xi_n^+ = \{\lambda \in \Xi_n : \mathscr{A} \subset \hat{A}_\lambda, \hat{A}_\lambda \ne \mathscr{A}\}$ (overfitted models). The proof then boils down to show that $P\Big(\inf_{\lambda\in\Xi_n^-}[\mathrm{BIC}^{\mathrm{H}}(\lambda) - \mathrm{BIC}^{\mathrm{H}}(\lambda_n)] > 0\Big) \to 1$ and $P\Big(\inf_{\lambda\in\Xi_n^+}[\mathrm{BIC}^{\mathrm{H}}(\lambda) - \mathrm{BIC}^{\mathrm{H}}(\lambda_n)] > 0\Big) \to 1$ as $n \to \infty$. Note that BIC contains both a goodness-of-fit part and a model complexity part. The basic idea is to show that for underfitted models, the goodness-of-fit part dominates, while for overfitted models, the model complexity part dominates.

Let $\widehat{Q}_n^\lambda = (nK)^{-1}\sum_{i=1}^n\sum_{k=1}^K \rho_{\tau_k}(y_i - \hat{\alpha}_k^\lambda - \mathbf{x}_i^{\mathrm{T}}\hat{\beta}^\lambda)$, where $(\widehat{\alpha}^\lambda, \widehat{\beta}^\lambda)$ is the two-step LLA estimator to the folded concave penalized CQR (6) with lasso initialization. Also, let $Q_n^* = (nK)^{-1}\sum_{i=1}^n\sum_{k=1}^K \rho_{\tau_k}(y_i - \alpha_k^* - \mathbf{x}_i^{\mathrm{T}}\beta^*)$. For any $A \subset \{1,2,\ldots,p\}$, let $(\widehat{\alpha}^A, \widehat{\beta}^A)$ be the estimator obtained by fitting the canonical CQR to model A, i.e.,

$$(\widehat{\alpha}^A, \widehat{\beta}^A) = \operatorname*{arg\,min}_{\alpha,\beta:\beta_{A^c}=\mathbf{0}} \frac{1}{nK}\sum_{i=1}^n\sum_{k=1}^K \rho_{\tau_k}(y_i - \alpha_k - \mathbf{x}_i^{\mathrm{T}}\beta). \tag{19}$$

Define $\widehat{Q}_n^A = (nK)^{-1}\sum_{i=1}^n\sum_{k=1}^K \rho_{\tau_k}(y_i - \hat{\alpha}_k^A - \mathbf{x}_i^{\mathrm{T}}\hat{\beta}^A)$. For any $\lambda \in \Xi_n$, recall that $\hat{A}_\lambda = \{1 \le j \le p : \hat{\beta}_j^\lambda \ne 0\}$ corresponds to the active set of the two-step LLA estimator $(\widehat{\alpha}^\lambda, \widehat{\beta}^\lambda)$. By optimality of $(\widehat{\alpha}^{\hat{A}_\lambda}, \widehat{\beta}^{\hat{A}_\lambda})$ in (39), we have $\widehat{Q}_n^{\hat{A}_\lambda} \le \widehat{Q}_n^\lambda$.

Let $\mathscr{G}_n^+ = \{A : A \supset \mathscr{A}, A \ne \mathscr{A}, |A| \le J_n\}$. It can be seen that $\{\hat{A}_\lambda : \lambda \in \Xi_n^+\} \subset \mathscr{G}_n^+$. For $r > 0$ and $A \in \mathscr{G}_n^+$, let $B_A(r) = \{(\delta,\Delta) \in \mathbb{R}^K \times \mathbb{R}^p : \|\delta\|^2 + \|\Delta_A\|^2 \le r^2, \Delta_{A^c} = \mathbf{0}\}$ and $S_A(r) = \{(\delta,\Delta) \in \mathbb{R}^K \times \mathbb{R}^p : \|\delta\|^2 + \|\Delta_A\|^2 = r^2, \Delta_{A^c} = \mathbf{0}\}$.

Case I: overfitted models. By Theorem 2, under the assumptions of this theorem, we have

$$P(\hat{A}_{\lambda_n} \ne \mathscr{A}) \to \mathscr{O}(1) \quad \text{as } n \to \infty.$$

Therefore, for any $\lambda \in \Xi_n^+$, we have

$$\begin{aligned}
&\Pr\Big(\inf_{\lambda\in\Xi_n^+}[\mathrm{BIC}^{\mathrm{H}}(\lambda) - \mathrm{BIC}^{\mathrm{H}}(\lambda_{\mathrm{n}})] > 0\Big) \\
&= \Pr\Big(\inf_{\lambda\in\Xi_n^+}[\mathrm{BIC}^{\mathrm{H}}(\lambda) - \mathrm{BIC}^{\mathrm{H}}(\lambda_{\mathrm{n}})] > 0, \hat{A}_{\lambda_n} = \mathscr{A}\Big) \\
&\quad + \Pr\Big(\inf_{\lambda\in\Xi_n^+}[\mathrm{BIC}^{\mathrm{H}}(\lambda) - \mathrm{BIC}^{\mathrm{H}}(\lambda_{\mathrm{n}})] > 0, \hat{A}_{\lambda_n} \ne \mathscr{A}\Big) \\
&= \Pr\Big(\inf_{\lambda\in\Xi_n^+}\Big[(\widehat{Q}_n^\lambda - \widehat{Q}_n^{\mathscr{A}}) + (|\hat{A}_\lambda| - s)\frac{C_n\log(p)}{n}\Big] > 0\Big) + \mathscr{O}(1) \\
&\ge \Pr\Big(\inf_{\lambda\in\Xi_n^+}\Big[(\widehat{Q}_n^{\hat{A}_\lambda} - \widehat{Q}_n^{\mathscr{A}}) + (|\hat{A}_\lambda| - s)\frac{C_n\log(p)}{n}\Big] > 0\Big) + \mathscr{O}(1),
\end{aligned}$$

where the last inequality follows from the fact that $\widehat{Q}_n^{\lambda} \geq \widehat{Q}_n^{\hat{A}_\lambda}$. Moreover, note that $\widehat{Q}_n^{\hat{A}_\lambda} \leq \widehat{Q}_n^{\mathscr{A}} \leq Q_n^*$ due to inclusion $\mathscr{A} \subset \hat{A}_\lambda$.

Let $F(\delta,\Delta) = Q_n(\alpha^*+\delta,\beta^*+\Delta) - Q_n(\alpha^*,\beta^*)$ for $(\delta,\Delta) \in \mathbb{R}^K \times \mathbb{R}^p$. For each model $A \in \mathscr{G}_n^+$, let $r_A^* = 16K(M + \bar{\zeta}^{1/2})(\underline{f}\underline{\zeta})^{-1}\sqrt{(|A|+1)\log(p)/n}$. If we can show that $\inf_{(\delta,\Delta)\in S_A(r_A^*)} F(\delta,\Delta) > 0$, then by convexity of $\rho_\tau(\cdot)$, we must have $\|\hat{\alpha}^A - \alpha^*\|_2^2 + \|\hat{\beta}^A - \beta^*\|_2^2 \leq (r_A^*)^2$. Indeed, by Knight's identity (see (28) of the appendix) and the mean value theorem, we have

$$\begin{aligned}
&\inf_{(\delta,\Delta)\in S_A(r_A^*)} \mathbb{E}[F(\delta,\Delta)] \\
&= \inf_{(\delta,\Delta)\in S_A(r_A^*)} \frac{1}{nK}\sum_{i=1}^n\sum_{k=1}^K \int_0^{\delta_k + \mathbf{x}_i^{\mathrm{T}}\Delta} [F(\alpha_k^* + t) - F(\alpha_k^*)]\mathrm{d}t \\
&= \inf_{(\delta,\Delta)\in S_A(r_A^*)} \frac{1}{nK}\sum_{i=1}^n\sum_{k=1}^K \int_0^{\delta_k + \mathbf{x}_i^{\mathrm{T}}\Delta} [tf(\alpha_k^* + \bar{u}_{ik,t})]\mathrm{d}t.
\end{aligned} \tag{20}$$

Note that J_n satisfies $J_n^2\log(p)/n = \mathscr{O}(1)$. Therefore, for any $1 \leq i \leq n, 1 \leq k \leq K$ and $(\delta,\Delta) \in S_A(r_A^*)$, we have

$$\begin{aligned}
&|\delta_k + \mathbf{x}_i^{\mathrm{T}}\Delta| \leq \sqrt{1 + \|\mathbf{x}_{iA}\|^2}\cdot\sqrt{\delta_k^2 + \|\Delta_A\|_2^2} \\
&\leq Mr_A^*\sqrt{|A|+1} = 16KM(M+\bar{\zeta}^{1/2})(\underline{f}\underline{\zeta})^{-1}(|A|+1)\sqrt{\log(p)/n} \\
&\leq 16KM(M+\bar{\zeta}^{1/2})(\underline{f}\underline{\zeta})^{-1}(J_n+1)\sqrt{\log(p)/n} = \mathscr{O}(\mathscr{U}_0).
\end{aligned}$$

It follows from condition (C0) and (40) that

$$\inf_{(\delta,\Delta)\in S_A(r_A^*)} \mathbb{E}[F(\delta,\Delta)] \geq \inf_{(\delta,\Delta)\in S_A(r_A^*)} \frac{\underline{f}}{2nK}\sum_{k=1}^K\sum_{i=1}^n (\delta_k + \mathbf{x}_i^{\mathrm{T}}\Delta)^2 \geq \frac{\underline{f}}{2K}\underline{\zeta}(r_A^*)^2.$$

Therefore, by Lemma 9, with probability at least $1 - \exp\{-nt^2/[32\bar{\zeta}(r_A^*)^2]\}$, we have

$$\begin{aligned}
\inf_{(\delta,\Delta)\in S_A(r_A^*)} F(\delta,\Delta) &\geq \inf_{(\delta,\Delta)\in S_A(r_A^*)} \mathbb{E}[F(\delta,\Delta)] - z(A,r_A^*) \\
&\geq \frac{\underline{f}}{2K}\underline{\zeta}(r_A^*)^2 - 4Mr_A^*\sqrt{\frac{|A|+1}{n}} - t.
\end{aligned}$$

Now take $t = 8r_A^*\sqrt{\bar{\zeta}(|A|+1)\log(p)/n}$. It follows that for each $A \in \mathscr{G}_n^+$, with probability at least $b_n^A = 1 - \exp\{-2(|A|+1)\log(p)\}$, we have

$$\inf_{(\delta,\Delta)\in S_A(r_A^*)} F(\delta,\Delta) \geq \frac{\underline{f}\underline{\zeta}(r_A^*)^2}{2K} - r_A^*\sqrt{\frac{|A|+1}{n}}\left(8\sqrt{\bar{\zeta}\log(p)} + 4M\right) > 0,$$

which immediately implies that

$$\|\hat{\alpha}^A - \alpha\|_2^2 + \|\hat{\beta}^A - \beta^*\|_2^2 \leq (r_A^*)^2.$$

Now by the Bonferroni inequality, we have $\|\hat{\alpha}^A-\alpha\|_2^2+\|\hat{\beta}^A-\beta^*\|_2^2\le (r_A^*)^2$ for all $A\in\mathscr{G}_n^+$ simultaneously with probability at least

$$\begin{aligned} b_n &= 1-\sum_{|A|=s+1}^{J_n}\binom{p-s}{|A|-s}(1-b_n^A)\\ &= 1-\sum_{k=1}^{J_n-s}\binom{p-s}{k}\exp\{-2(k+s+1)\log(p)\}\\ &\ge 1-p^{-2(s+1)}\sum_{k=1}^{p-s}\binom{p-s}{k}\Big(\frac{1}{p^2}\Big)^k\\ &= 1-p^{-2(s+1)}\Big[\Big(1+\frac{1}{p^2}\Big)^{p-s}-1\Big]\to 1 \quad \text{as } p\to\infty. \end{aligned} \tag{21}$$

Now we derive the upper bound for $\sup_{A\in\mathscr{G}_n^+}|\widehat{Q}_n^A-Q_n^*|$. Let $\hat{\delta}^A=\hat{\alpha}^A-\alpha^*$ and $\hat{\Delta}^A=\hat{\beta}^A-\beta^*$. Observe that

$$\begin{aligned} |\widehat{Q}_n^A-Q_n^*| &= \Big|\frac{1}{nK}\sum_{i=1}^n\sum_{k=1}^K\{\rho_{\tau_k}(y_i-\hat{\alpha}_k^A-\mathbf{x}_i^{\mathrm{T}}\hat{\beta}^A)-\rho_{\tau_k}(y_i-\alpha_k^*-\mathbf{x}_i^{\mathrm{T}}\beta^*)\}\Big|\\ &\le \big|\mathbb{E}\{F(\hat{\delta}^A,\hat{\beta}^A)\}\big|+z(A,r_A^*). \end{aligned}$$

Similarly, we have

$$\begin{aligned} \big|\mathbb{E}\{F(\hat{\delta}^A,\hat{\beta}^A)\}\big| &= \Big|\frac{1}{nK}\sum_{i=1}^n\sum_{k=1}^K\int_0^{\hat{\delta}_k^A+\mathbf{x}_i^{\mathrm{T}}\hat{\Delta}^A}[F(\alpha_k^*+t)-F(\alpha_k^*)]\mathrm{d}t\Big|\\ &\le \frac{1}{nK}\sum_{i=1}^n\sum_{k=1}^K\int_0^{|\hat{\delta}_k^A+\mathbf{x}_i^{\mathrm{T}}\hat{\Delta}^A|}[tf(\alpha_k^*+\bar{u}_{ik,t})]\mathrm{d}t\\ &\le \frac{\bar{f}}{2nK}\sum_{i=1}^n\sum_{k=1}^K(\hat{\delta}^A+\mathbf{x}_i^{\mathrm{T}}\hat{\Delta}^A)^2\\ &\le \frac{1}{2}\bar{f}\bar{\zeta}(\|\hat{\delta}^A\|_2^2+\|\hat{\Delta}^A\|_2^2)\le\frac{1}{2}\bar{f}\bar{\zeta}(r_A^*)^2. \end{aligned}$$

It follows that with probability at least b_n,

$$\begin{aligned} |\widehat{Q}_n^A-Q_n^*| &\le \frac{1}{2}\bar{f}\bar{\zeta}(r_A^*)^2+4Mr_A^*\sqrt{\frac{|A|+1}{n}}+8r_A^*\sqrt{\frac{\bar{\zeta}(|A|+1)\log(p)}{n}}\\ &\le \frac{128K^2(M+\bar{\zeta}^{1/2})^2}{\underline{f}\,\underline{\zeta}}\,\frac{(|A|+1)\log(p)}{n} \end{aligned}$$

holds for all $A\in\mathscr{G}_n^+$. Now going back to BIC, we have

$$\begin{aligned} &\Pr\Big(\inf_{\lambda\in\Xi_n^+}\Big[(\widehat{Q}_n^{\hat{A}_\lambda}-\widehat{Q}_n^{\mathscr{A}})+(|\hat{A}_\lambda|-s)\frac{C_n\log(p)}{n}\Big]>0\Big)\\ &\ge \Pr\Big(\frac{C_n\log(p)}{n}-\sup_{A\in\mathscr{G}_n^+}\frac{\widehat{Q}_n^{\mathscr{A}}-\widehat{Q}_n^A}{(|A|-s)}>0\Big). \end{aligned}$$

Therefore, with probability at least b_n, we have

$$\sup_{A\in\mathscr{G}_n^+}\frac{\widehat{Q}_n^{\mathscr{A}}-\widehat{Q}_n^A}{(|A|-s)}\le\sup_{A\in\mathscr{G}_n^+}\frac{\widehat{Q}_n^*-\widehat{Q}_n^A}{(|A|-s)}=\mathscr{O}_P\Big(\frac{s\log(p)}{n}\Big).$$

Since $s=\mathscr{O}(1)$ and C_n diverges with n, we have $s\log(p)/n=o(C_n\log(p)/n)$. It follows that

$$\Pr\Big(\frac{C_n\log(p)}{n}-\sup_{A\in\mathscr{G}_n^+}\frac{\widehat{Q}_n^{\mathscr{A}}-\widehat{Q}_n^A}{(|A|-s)}>0\Big)\to 1 \quad \text{as } n\to\infty,$$

which implies that $\Pr\big(\inf_{\lambda\in\Xi_n^+}[\mathrm{BIC}^{\mathrm{H}}(\lambda)-\mathrm{BIC}^{\mathrm{H}}(\lambda_{\mathrm{n}})]>0\big)\to 1$ as $n\to\infty$.

Case II: underfitted models. For any $\lambda \in \Xi_n^-$, similar to Case I, we have

$$\begin{aligned}&\Pr\Big(\inf_{\lambda\in\Xi_n^-}[\mathrm{BIC}^{\mathrm{H}}(\lambda)-\mathrm{BIC}^{\mathrm{H}}(\lambda_{\mathrm{n}})]>0\Big)\\ &\geq \Pr\bigg(\inf_{\lambda\in\Xi_n^-}\Big[(\widehat{Q}_n^{\hat{A}_\lambda}-\widehat{Q}_n^{\mathscr{A}})+(|\hat{A}_\lambda|-s)\frac{C_n\log(p)}{n}\Big]>0\bigg)+\mathscr{O}(1).\end{aligned}$$

Define $\mathrm{BIC}^{\mathrm{H}}(A)=\widehat{Q}_n^A+|A|C_n\log(p)/n$ and let $\mathscr{G}_n^-=\{A:|A|\leq J_n,\mathscr{A}\not\subset A\}$. We can see that $\{\hat{A}_\lambda:\lambda\in\Xi_n^-\}\subset\mathscr{G}_n^-$. It suffices to show $\inf_{A\in\mathscr{G}_n^-}\mathrm{BIC}^{\mathrm{H}}(A)>\mathrm{BIC}^{\mathrm{H}}(\mathscr{A})$ with probability tending to one as $n\to\infty$. For any $A\in\mathscr{G}_n^-$, let $\bar{A}=A\cup\mathscr{A}$. Let $\theta=\min_{j\in\mathscr{A}}|\beta_j^*|$. Since $A\not\supset\mathscr{A}$, we must have $\|\hat{\alpha}^A-\alpha^*\|_2^2+\|\hat{\beta}^A-\beta^*\|_2\geq\theta^2$. However, since $\bar{A}\supset\mathscr{A}$ and $|\bar{A}|\leq 2J_n$, using Lemma 9, we can similarly show as in Case I that $\|\hat{\alpha}^{\bar{A}}-\alpha^*\|_2^2+\|\hat{\beta}^{\bar{A}}-\beta^*\|_2^2\leq\theta^2$ with probability at least $b_n^{\bar{A}}=1-\exp\{-2(|\bar{A}|+1)\log(p)\}$ as long as

$$\theta>8K(\underline{f}\underline{\zeta})^{-1}(2\sqrt{\bar{\zeta}\log(p)}+M)\sqrt{\frac{2J_n+1}{n}},$$

which is implied by the assumption $\sqrt{J_n\log(p)/n}=\mathscr{O}(\theta)$. It then follows that $\|\hat{\alpha}^{\bar{A}}-\alpha^*\|_2^2+\|\hat{\beta}^{\bar{A}}-\beta^*\|_2^2\leq\theta^2$ holds for all $A\in\mathscr{G}_n^-$ with probability at least $\tilde{b}_n\to 1$ as $n\to\infty$, where

$$\tilde{b}_n=1-\sum_{|\bar{A}=s+1}^{2J_n}\binom{p-s}{|\bar{A}|-s}(1-b_n^{\bar{A}})\geq 1-p^{-2(s+1)}\Big[\Big(1+\frac{1}{p^2}\Big)^{p-s}-1\Big].$$

Therefore, there exists $a\in[0,1]$, $\bar{\alpha}^{\bar{A}}=a\hat{\alpha}^A+(1-a)\hat{\alpha}^{\bar{A}}$ and $\bar{\beta}^{\bar{A}}=a\hat{\beta}^A+(1-a)\hat{\beta}^{\bar{A}}$ such that $\|\bar{\alpha}^{\bar{A}}-\alpha^*\|_2^2+\|\bar{\beta}^{\bar{A}}-\beta^*\|_2^2=\theta^2$. By convexity of ρ_τ and the fact that $\widehat{Q}_n^A\geq\widehat{Q}_n^{\bar{A}}$, we have $\bar{Q}_n^{\bar{A}}=(nK)^{-1}\sum_{i=1}^n\sum_{k=1}^K\rho_{\tau_k}(y_i-\bar{\alpha}_k^{\bar{A}}-\mathbf{x}_i^{\mathrm{T}}\bar{\beta}^{\bar{A}})\leq\widehat{Q}_n^A$. Note that $\widehat{Q}_n^{\bar{A}}\leq\widehat{Q}_n^{\mathscr{A}}\leq Q_n^*$. It follows that $\widehat{Q}_n^A-\widehat{Q}_n^{\bar{A}}\geq\bar{Q}_n^{\bar{A}}-Q_n^*$. For ease of notation, let $\bar{\delta}^{\bar{A}}=\bar{\alpha}^{\bar{A}}-\alpha^*$ and $\bar{\Delta}^{\bar{A}}=\bar{\beta}^{\bar{A}}-\beta^*$. It can be seen that

$$\bar{Q}_n^{\bar{A}}-Q_n^*\geq\mathbb{E}[F(\bar{\delta}^{\bar{A}},\bar{\Delta}^{\bar{A}})]-z(\bar{A},\theta).$$

Following similar arguments from Case I and noting that the support of $\bar{\beta}^{\bar{A}}$ is a subset of $\bar{A}$, we can show that with probability at least $\tilde{b}_n$, for all $A\in\mathscr{G}_n^-$, we have

$$\bar{Q}_n^{\bar{A}}-Q_n^*\geq\frac{\underline{f}}{2K}\underline{\zeta}\theta^2-4\theta\sqrt{\frac{|\bar{A}|+1}{n}}(2\sqrt{\bar{\zeta}\log(p)}+M).$$

Now we have

$$\begin{aligned}&\mathrm{BIC}^{\mathrm{H}}(A)-\mathrm{BIC}^{\mathrm{H}}(\bar{A})=(\widehat{Q}_n^A-\widehat{Q}_n^{\bar{A}})+(|A|-|\bar{A}|)\frac{C_n\log(p)}{n}\\ &\geq(\bar{Q}_n^{\bar{A}}-Q_n^*)-\frac{C_ns\log(p)}{n}.\end{aligned}$$

Since $\sqrt{C_ns\log(p)/n}=\mathscr{O}(\theta)$, it can be seen that with probability tending to one, we have $\inf_{A\in\mathscr{G}_n^-}\mathrm{BIC}^{\mathrm{H}}(A)-\mathrm{BIC}^{\mathrm{H}}(\bar{A})>0$. Following similar arguments as in Case I, we can show $\mathrm{BIC}^{\mathrm{H}}(\bar{A})\geq\inf_{S\supset\mathscr{A},|S|\leq 2J_n}\mathrm{BIC}^{\mathrm{H}}(S)\geq\mathrm{BIC}^{\mathrm{H}}(\mathscr{A})$ with probability tending to one. Case II then follows by noting that

$$\begin{aligned}&\inf_{A\in\mathscr{G}_n^-}[\mathrm{BIC}^{\mathrm{H}}(A)-\mathrm{BIC}^{\mathrm{H}}(\mathscr{A})]\\ &=\inf_{A\in\mathscr{G}_n^-}[\mathrm{BIC}^{\mathrm{H}}(A)-\mathrm{BIC}^{\mathrm{H}}(\bar{A})+\mathrm{BIC}^{\mathrm{H}}(\bar{A})-\mathrm{BIC}^{\mathrm{H}}(\mathscr{A})]\\ &\geq\inf_{A\in\mathscr{G}_n^-}[\mathrm{BIC}^{\mathrm{H}}(A)-\mathrm{BIC}^{\mathrm{H}}(\bar{A})].\end{aligned}$$

□

Proof of Theorem 4. See Section A of the appendix. □

Proof of Lemma 2. See Lemma 1 of [23]. □

APPENDIX A
PROOFS

Proof of Lemma 3. Let $\zeta = (\zeta_1, \ldots, \zeta_K)^{\mathrm{T}}$ and $\xi = (\xi_1, \ldots, \xi_p)^{\mathrm{T}}$, where

$$\zeta_k = -\frac{1}{nK}\sum_{i=1}^{n}\big[\tau_k - I(\varepsilon_i \le \alpha_k^*)\big],\ 1 \le k \le K,$$

and

$$\xi_j = -\frac{1}{nK}\sum_{k=1}^{K}\sum_{i=1}^{n}\big[\tau_k - I(\varepsilon_i \le \alpha_k^*)\big]x_{ij},\ 1 \le j \le p.$$

Note that $(\zeta^{\mathrm{T}}, \xi^{\mathrm{T}})^{\mathrm{T}} \in \partial Q_n(\alpha^*, \beta^*)$, where the subdifferential is taken with respect to α and β. By convexity of $Q_n(\alpha, \beta)$ and optimality of $(\widehat{\alpha}_\lambda, \widehat{\beta}_\lambda)$, we have

$$\begin{aligned}
0 &\ge Q_n(\widehat{\alpha}_\lambda, \widehat{\beta}_\lambda) - Q_n(\alpha^*, \beta^*) + \lambda(\|\widehat{\beta}_\lambda\|_1 - \|\beta^*\|_1) \\
&\ge \zeta^{\mathrm{T}}(\widehat{\alpha}_\lambda - \alpha^*) + \xi^{\mathrm{T}}(\widehat{\beta}_\lambda - \beta^*) + \lambda(\|\widehat{\beta}_\lambda\|_1 - \|\beta^*\|_1) \\
&\ge -\|\zeta\|_\infty \cdot \|\widehat{\alpha}_\lambda - \alpha^*\|_1 - \|\xi\|_\infty \cdot \|\widehat{\beta}_\lambda - \beta^*\|_1 \\
&\quad + \lambda\big(\|\widehat{\beta}_{\lambda,\mathscr{A}^{\mathrm{c}}} - \beta^*_{\mathscr{A}^{\mathrm{c}}}\|_1 - \|\widehat{\beta}_{\lambda,\mathscr{A}} - \beta^*_{\mathscr{A}}\|_1\big),
\end{aligned}$$

which implies that

$$\begin{aligned}
(\lambda - \|\xi\|_\infty)\|\widehat{\beta}_{\lambda,\mathscr{A}^{\mathrm{c}}} - \beta^*_{\mathscr{A}^{\mathrm{c}}}\|_1 &\le (\lambda + \|\xi\|_\infty)\|\widehat{\beta}_{\lambda,\mathscr{A}} - \beta^*_{\mathscr{A}}\|_1 \\
&\quad + \|\zeta\|_\infty \cdot \|\widehat{\alpha}_\lambda - \alpha^*\|_1.
\end{aligned} \tag{22}$$

Under event $\mathscr{E} = \{\|\zeta\|_\infty \le 3\lambda/(2K), \|\xi\|_\infty \le \lambda/2\}$, it follows from (22) that

$$\|\widehat{\Delta}^\lambda_{\mathscr{A}^{\mathrm{c}}}\|_1 \le 3\|\widehat{\Delta}^\lambda_{\mathscr{A}}\|_1 + \tfrac{3}{K}\|\widehat{\delta}^\lambda\|_1.$$

The lemma then follows from Hoeffding's inequality

$$\begin{aligned}
\Pr(\mathscr{E}) &\ge 1 - \Pr\Big(\|\zeta\|_\infty > \frac{3\lambda}{2K}\Big) - \Pr\Big(\|\xi\|_\infty > \frac{\lambda}{2}\Big) \\
&\ge 1 - \sum_{k=1}^{K}\Pr\Big(\Big|-\frac{1}{nK}\sum_{i=1}^{n}[\tau_k - I(\varepsilon_i \le \alpha_k^*)]\Big| > \frac{3\lambda}{2K}\Big) \\
&\quad - \sum_{j=1}^{p}\Pr\Big(\Big|-\frac{1}{nK}\sum_{i=1}^{n}x_{ij}\sum_{k=1}^{K}[\tau_k - I(\varepsilon_i \le \alpha_k^*)]\Big| > \frac{\lambda}{2}\Big) \\
&\ge 1 - 2K\exp\Big(-\frac{9n\lambda^2}{2}\Big) - 2p\exp\Big(-\frac{n\lambda^2}{2M_0}\Big).
\end{aligned}$$

This proves the lemma. □

Proof of Lemma 4. First, let us show that the check loss $\rho_\tau(\cdot)$ is Lipschitz continuous with Lipschitz constant $\max(\tau, 1-\tau)$. To see it, note that for any $u_1, u_2 \in \mathbb{R}$, we have

$$\begin{aligned}
|\rho_\tau(u_1) - \rho_\tau(u_2)| &= |(\tau - 0.5)(u_1 - u_2) + 0.5(|u_1| - |u_2|)| \\
&\le (|\tau - 0.5| + 0.5)|u_1 - u_2| = \max(\tau, 1-\tau)|u_1 - u_2|.
\end{aligned}$$

Now let $\delta = \alpha - \alpha^*$, $\Delta = \beta - \beta^*$, and define

$$\begin{aligned} U_i(\delta,\Delta) &= \frac{1}{K}\sum_{k=1}^{K}\rho_{\tau_k}(y_i - \alpha_k - \mathbf{x}_i^{\mathrm{T}}\beta) - \frac{1}{K}\sum_{k=1}^{K}\rho_{\tau_k}(y_i - \alpha_k^* - \mathbf{x}_i^{\mathrm{T}}\beta^*) \\ &= \frac{1}{K}\sum_{k=1}^{K}\rho_{\tau_k}(r_{ik}^* - \delta_k - \mathbf{x}_i^{\mathrm{T}}\Delta) - \frac{1}{K}\sum_{k=1}^{K}\rho_{\tau_k}(r_{ik}^*), \end{aligned}$$

where $r_{ik}^* = y_i - \alpha_k^* - \mathbf{x}_i^{\mathrm{T}}\beta^* = \varepsilon_i - \alpha_k^*$, $1 \le i \le n$, $1 \le k \le K$. It follows immediately that

$$e(r) = \sup_{(\delta,\Delta)\in\mathscr{C}_r}\left|\frac{1}{n}\sum_{i=1}^{n}\big[U_i(\delta,\Delta) - \mathbb{E}U_i(\delta,\Delta)\big]\right|.$$

By Lipschitz continuity of the check loss, it follows that

$$\begin{aligned} |U_i(\delta,\Delta)| &\le \frac{1}{K}\sum_{k=1}^{K}|\rho_{\tau_k}(r_{ik}^* - \delta_k - \mathbf{x}_i^{\mathrm{T}}\Delta) - \rho_{\tau_k}(r_{ik}^*)| \\ &\le \frac{1}{K}\sum_{k=1}^{K}\max(\tau_k, 1-\tau_k)|\delta_k + \mathbf{x}_i^{\mathrm{T}}\Delta| \le \frac{1}{K}\sum_{k=1}^{K}|\delta_k + \mathbf{x}_i^{\mathrm{T}}\Delta|,\ 1 \le i \le n. \end{aligned} \tag{23}$$

Now applying Massart's concentration inequality (Theorem 14.2, [25]), we obtain

$$\Pr(e(r) \ge \mathbb{E}[e(r)] + t) \le \exp\left(-\frac{n^2t^2}{8b_n^2(r)}\right), \tag{24}$$

where $b_n^2(r) = \sup_{(\delta,\Delta)\in\mathscr{C}_r}\sum_{i=1}^{n}\mathrm{var}(U_i(\delta,\Delta))$. First, we derive the upper bound on $b_n^2(r)$. Note that by (23) and Cauchy–Schwarz inequality

$$\begin{aligned} b_n^2(r) &= \sup_{(\delta,\Delta)\in\mathscr{C}_r}\sum_{i=1}^{n}\mathbb{E}\big[U_i(\delta,\Delta) - \mathbb{E}U_i(\delta,\Delta)\big]^2 \le 4\sup_{(\delta,\Delta)\in\mathscr{C}_r}\sum_{i=1}^{n}\left[\sum_{k=1}^{K}\frac{1}{K}|\delta_k + \mathbf{x}_i^{\mathrm{T}}\Delta|\right]^2 \\ &\le 4\sup_{(\delta,\Delta)\in\mathscr{C}_r}\sum_{i=1}^{n}\left(\sum_{k=1}^{K}\frac{1}{K}\right)\left[\sum_{k=1}^{K}\frac{1}{K}(\delta_k + \mathbf{x}_i^{\mathrm{T}}\Delta)^2\right] \le 4nr^2. \end{aligned}$$

Next, we derive the upper bound on $\mathbb{E}\big[e(r)\big]$. By applying the symmetrization procedure [26] and the contraction principle [27], we have

$$\begin{aligned} \mathbb{E}[e(r)] &\le 2\mathbb{E}\left[\sup_{(\delta,\Delta)\in\mathscr{C}_r}\frac{1}{n}\left|\sum_{i=1}^{n}\xi_i U_i(\delta,\Delta)\right|\right] \\ &\le \frac{2}{nK}\sum_{k=1}^{K}\mathbb{E}\left[\sup_{(\delta,\Delta)\in\mathscr{C}_r}\left|\sum_{i=1}^{n}\xi_i\{\rho_{\tau_k}(r_{ik}^* - \delta_k - \mathbf{x}_i^{\mathrm{T}}\Delta) - \rho_{\tau_k}(r_{ik}^*)\}\right|\right] \\ &\le \frac{4}{nK}\sum_{k=1}^{K}\mathbb{E}\left[\sup_{(\delta,\Delta)\in\mathscr{C}_r}\left|\sum_{i=1}^{n}\xi_i(\delta_k + \mathbf{x}_i^{\mathrm{T}}\Delta)\right|\right], \end{aligned} \tag{25}$$

where $\xi_1, \ldots, \xi_n$ are i.i.d. Rademacher random variables that satisfy $\Pr(\xi_i = -1) = \Pr(\xi_i = 1) = 1/2$ and that are independent of $\varepsilon_1, \ldots, \varepsilon_n$.

For $(\delta,\Delta) \in \mathscr{C}_r$, by condition (C1) and the Cauchy–Schwarz inequality, we have

$$r^2 \ge \frac{\kappa_0}{K}(\|\delta\|_2^2 + K\|\Delta_{\mathscr{A}}\|_2^2) \ge \frac{\kappa_0}{K^2}\|\delta\|_1^2 + \frac{\kappa_0}{s}\|\Delta_{\mathscr{A}}\|_1^2, \tag{26}$$

which implies that $\|\delta\|_1 \le rK/\sqrt{\kappa_0}$ and $\|\Delta_{\mathscr{A}}\|_1 \le r\sqrt{s/\kappa_0}$. Now let $\xi = (\xi_1, \ldots, \xi_n)^{\mathrm{T}}$. Note that for any $t \in \mathbb{R}$, we have by Taylor expansion

$$\begin{aligned} \mathbb{E}\big[\exp(tX_j^{\mathrm{T}}\xi)\big] &= \prod_{i=1}^{n}\left[\frac{1}{2}(e^{tx_{ij}} + e^{-tx_{ij}})\right] \\ &\le \prod_{i=1}^{n}\exp\Big(\frac{1}{2}t^2x_{ij}^2\Big) = \exp\left(\frac{t^2}{2}\sum_{i=1}^{n}x_{ij}^2\right),\ 0 \le j \le p. \end{aligned}$$

Letting $t>0$, by Jensen's inequality, we have

$$\begin{aligned}
&\exp\big(t\mathbb{E}\big[\|\mathbb{X}^{\mathrm{T}}\xi\|_\infty\big]\big)=\exp\Big(t\mathbb{E}\max_{0\le j\le p}|X_j^{\mathrm{T}}\xi|\Big)\le\mathbb{E}\exp\Big(t\max_{0\le j\le p}|X_j^{\mathrm{T}}\xi|\Big)\\
&=\mathbb{E}\Big[\max_{0\le j\le p}\exp(t|X_j^{\mathrm{T}}\xi|)\Big]\le\mathbb{E}\max_{0\le j\le p}\Big(e^{tX_j^{\mathrm{T}}\xi}+e^{-tX_j^{\mathrm{T}}\xi}\Big)\\
&\le\sum_{j=0}^{p}\mathbb{E}\Big(e^{tX_j^{\mathrm{T}}\xi}+e^{-tX_j^{\mathrm{T}}\xi}\Big)\le 2\sum_{j=0}^{p}\exp\Big(\frac{t^2}{2}\|X_j\|_2^2\Big)\\
&\le 2(1+p)\exp\Big(\frac{t^2}{2}\max_{0\le j\le p}\|X_j\|_2^2\Big)=2(1+p)\exp\Big(\frac{1}{2}nM_0t^2\Big),
\end{aligned}$$

which implies that

$$\mathbb{E}\big(\|\mathbb{X}^{\mathrm{T}}\xi\|_\infty\big)\le\frac{1}{t}\big[\log 2+\log(1+p)\big]+\frac{nM_0}{2}t,\ t>0.$$

Taking $t=\sqrt{2[\log 2+\log(1+p)]/(nM_0)}$ and noting that $p\ge 3$ by condition (C0), we obtain

$$\mathbb{E}\big(\|\mathbb{X}^{\mathrm{T}}\xi\|_\infty\big)\le\sqrt{2nM_0[\log 2+\log(1+p)]}\le\sqrt{2M_0}\cdot\sqrt{n(1+\log p)}. \tag{27}$$

It then follows from (25), (27) and Hölder's inequality that

$$\begin{aligned}
\mathbb{E}[e(r)]&\le\frac{4}{nK}\mathbb{E}\big(\|\mathbb{X}^{\mathrm{T}}\xi\|_\infty\big)\cdot\sup_{(\delta,\Delta)\in\mathscr{C}_r}\sum_{k=1}^{K}\big(|\delta_k|+\|\Delta\|_1\big)\\
&\le\frac{4\sqrt{2M_0}}{n}\sqrt{n(1+\log p)}\sup_{(\delta,\Delta)\in\mathscr{C}_r}\big(K^{-1}\|\delta\|_1+\|\Delta\|_1\big)\\
&\le\frac{4\sqrt{2M_0}}{n}\sqrt{n(1+\log p)}\sup_{(\delta,\Delta)\in\mathscr{C}_r}\big[4K^{-1}\|\delta\|_1+4\|\Delta_{\mathscr{A}}\|_1\big]\\
&\le 16\sqrt{\frac{2M_0}{\kappa_0}}\sqrt{\frac{1+\log p}{n}}(\sqrt{s}+1)r.
\end{aligned}$$

The lemma then follows from (24). □

Proof of Lemma 5. By Knight's identity [2], we have for any two scalars $r\neq 0$ and s,

$$|r-s|-|r|=-s\big[I(r>0)-I(r<0)\big]+2\int_0^s\big[I(r\le t)-I(r\le 0)\big]\mathrm{d}t.$$

It follows that for any $\tau\in(0,1)$, when $r\neq 0$,

$$\begin{aligned}
&\rho_\tau(r-s)-\rho_\tau(r)=(\tau-0.5)\big[(r-s)-r\big]+0.5\big[|r-s|-|r|\big]\\
&=(0.5-\tau)s-0.5s\big[I(r>0)-I(r<0)\big]+\int_0^s\big[I(r\le t)-I(r\le 0)\big]\mathrm{d}t\\
&=s\big[I(r<0)-\tau\big]+\int_0^s\big[I(r\le t)-I(r\le 0)\big]\mathrm{d}t.
\end{aligned} \tag{28}$$

Let $r_{ik}^* = y_i - \alpha_k^* - \mathbf{x}_i^{\mathrm{T}}\beta^* = \varepsilon_i - \alpha_k^*$, $1 \le i \le n$, $1 \le k \le K$. Recall that ε has a density with respect to the Lebesgue measure. By condition (C0), identity (28) and the mean value theorem, we have for some $\bar{u}_{ik,t}$ between 0 and t,

$$\begin{aligned}
&\mathbb{E}\big[Q_n(\alpha^*+\delta,\beta^*+\Delta) - Q_n(\alpha^*,\beta^*)\big] \\
&= \frac{1}{nK}\sum_{i=1}^{n}\sum_{k=1}^{K}\int_0^{\delta_k+\mathbf{x}_i^{\mathrm{T}}\Delta}\big[F(\alpha_k^*+t) - F(\alpha_k^*)\big]\mathrm{d}t \\
&= \frac{1}{nK}\sum_{i=1}^{n}\sum_{k=1}^{K}\int_0^{\delta_k+\mathbf{x}_i^{\mathrm{T}}\Delta}\Big[tf(\alpha_k^*) + \frac{t^2}{2}f'(\alpha_k^*+\bar{u}_{ik,t})\Big]\mathrm{d}t \\
&\ge \frac{\underline{f}}{2nK}\sum_{i=1}^{n}\sum_{k=1}^{K}(\delta_k+\mathbf{x}_i^{\mathrm{T}}\Delta)^2 - \frac{\bar{f}'}{6nK}\sum_{i=1}^{n}\sum_{k=1}^{K}|\delta_k+\mathbf{x}_i^{\mathrm{T}}\Delta|^3.
\end{aligned} \tag{29}$$

For $(\delta,\Delta)\in\mathscr{C}$, note that if

$$\Big[\frac{1}{nK}\sum_{i=1}^{n}\sum_{k=1}^{K}(\delta_k+\mathbf{x}_i^{\mathrm{T}}\Delta)^2\Big]^{1/2} \le \frac{4q}{K^{1/2}\underline{f}^{1/2}}, \tag{30}$$

then by condition (C2), this implies that

$$\frac{\bar{f}'}{6nK}\sum_{i=1}^{n}\sum_{k=1}^{K}|\delta_k+\mathbf{x}_i^{\mathrm{T}}\Delta|^3 \le \frac{\underline{f}}{4nK}\sum_{i=1}^{n}\sum_{k=1}^{K}(\delta_k+\mathbf{x}_i^{\mathrm{T}}\Delta)^2,$$

which, together with (29), implies that for all $(\delta,\Delta)\in\mathscr{C}_{4q(K\underline{f})^{-1/2}}$,

$$\mathbb{E}\big[Q_n(\alpha^*+\delta,\beta^*+\Delta) - Q_n(\alpha^*,\beta^*)\big] \ge \frac{\underline{f}}{4nK}\sum_{i=1}^{n}\sum_{k=1}^{K}(\delta_k+\mathbf{x}_i^{\mathrm{T}}\Delta)^2.$$

To show that the lemma holds for all $(\delta,\Delta)\in\mathscr{C}$, define

$$\begin{aligned}
r_{\mathscr{C}} = \sup_{r>0}\Big\{ r\colon\ &\mathbb{E}\big[Q_n(\alpha^*+\delta,\beta^*+\Delta) - Q_n(\alpha^*,\beta^*)\big] \\
&\ge \frac{\underline{f}}{4nK}\sum_{i=1}^{n}\sum_{k=1}^{K}(\delta_k+\mathbf{x}_i^{\mathrm{T}}\Delta)^2,\ \forall(\delta,\Delta)\in\mathscr{C}_r\Big\}.
\end{aligned}$$

By previous arguments, we must have $r_{\mathscr{C}} \ge 4q(K\underline{f})^{-1/2}$. Now for any $(\delta,\Delta)\in\mathscr{C}$, let $r^2 = (nK)^{-1}\sum_{i=1}^{n}\sum_{k=1}^{K}(\delta_k + \mathbf{x}_i^{\mathrm{T}}\Delta)^2$. If $r < r_{\mathscr{C}}$, then by the definition of $r_{\mathscr{C}}$, we have

$$\mathbb{E}\big[Q_n(\alpha^*+\delta,\beta^*+\Delta) - Q_n(\alpha^*,\beta^*)\big] \ge \frac{\underline{f}}{4nK}\sum_{i=1}^{n}\sum_{k=1}^{K}(\delta_k+\mathbf{x}_i^{\mathrm{T}}\Delta)^2. \tag{31}$$

If instead $r \ge r_{\mathscr{C}}$, let $\delta' = r_{\mathscr{C}}\delta/r$ and $\Delta' = r_{\mathscr{C}}\Delta/r$. It can be seen immediately that $(nK)^{-1}\sum_{i=1}^{n}\sum_{k=1}^{K}(\delta_k'+\mathbf{x}_i^{\mathrm{T}}\Delta')^2 = r_{\mathscr{C}}^2$. By convexity of Q_n, we have

$$\begin{aligned}
&\mathbb{E}\big[Q_n(\alpha^*+\delta,\beta^*+\Delta) - Q_n(\alpha^*,\beta^*)\big] \\
&\ge \frac{r}{r_{\mathscr{C}}}\mathbb{E}\big[Q_n(\alpha^*+\delta',\beta^*+\Delta') - Q_n(\alpha^*,\beta^*)\big] \\
&\ge \frac{r}{r_{\mathscr{C}}}\frac{\underline{f}}{4}r_{\mathscr{C}}^2 \ge q\Big(\frac{\underline{f}}{K}\Big)^{1/2}\Big[\frac{1}{nK}\sum_{i=1}^{n}\sum_{k=1}^{K}(\delta_k+\mathbf{x}_i^{\mathrm{T}}\Delta)^2\Big]^{1/2}.
\end{aligned} \tag{32}$$

The lemma then follows from (31) and (32). □

Proof of Lemma 6. As with the proof of Lemma 4, define

$$U_i(\delta,\Delta) = \frac{1}{K}\sum_{k=1}^{K}\rho_{\tau_k}(r_{ik}^* - \delta_k - \mathbf{x}_i^{\mathrm{T}}\Delta) - \frac{1}{K}\sum_{k=1}^{K}\rho_{\tau_k}(r_{ik}^*),$$

where $r_{ik}^* = y_i - \alpha_k^* - \mathbf{x}_i^{\mathrm{T}}\beta^* = \varepsilon_i - \alpha_k^*,\ 1 \le i \le n,\ 1 \le k \le K$. Now applying Massart's concentration inequality, we get

$$\Pr(z(r) \ge \mathbb{E}[z(r)] + t) \le \exp\bigg(-\frac{n^2t^2}{8b_n^2(r)}\bigg), \tag{33}$$

where $b_n^2(r) = \sup_{(\delta,\Delta)\in B_{\mathscr{A}}(r)} \sum_{i=1}^n \mathrm{var}(U_i(\delta,\Delta))$. For ease of notation, let $\Delta_{\mathscr{A}}^k = (\delta_k, \Delta_{\mathscr{A}}^{\mathrm{T}})^{\mathrm{T}},\ 1 \le k \le K$. It follows from Lipschitz continuity of the check loss that

$$\begin{aligned} b_n^2(r) &\le \frac{4}{K} \sup_{(\delta,\Delta)\in B_{\mathscr{A}}(r)} \sum_{k=1}^K \sum_{i=1}^n (\delta_k + \mathbf{x}_i^{\mathrm{T}}\Delta)^2 \\ &= \frac{4}{K} \sup_{(\delta,\Delta)\in B_{\mathscr{A}}(r)} \sum_{k=1}^K (\Delta_{\mathscr{A}}^k)^{\mathrm{T}} \mathbb{X}_{\mathscr{A}_0}^{\mathrm{T}} \mathbb{X}_{\mathscr{A}_0} \Delta_{\mathscr{A}}^k \\ &\le \frac{4n}{K} \sup_{(\delta,\Delta)\in B_{\mathscr{A}}(r)} \sum_{k=1}^K \overline{\mu}\big[\delta_k^2 + \|\Delta_{\mathscr{A}}\|_2^2\big] \le 4n\overline{\mu}r^2. \end{aligned}$$

Moreover, by the symmetrization procedure and the contraction principle, we obtain

$$\begin{aligned} \mathbb{E}[z(r)] &\le \frac{4}{nK} \sum_{k=1}^K \mathbb{E}\bigg[\sup_{(\delta,\Delta)\in B_{\mathscr{A}}(r)} \bigg|\sum_{i=1}^n \xi_i(\delta_k + \mathbf{x}_{\mathscr{A}}^{\mathrm{T}}\Delta_{\mathscr{A}})\bigg|\bigg] \\ &\le \frac{4}{nK}\mathbb{E}\big(\|\mathbb{X}_{\mathscr{A}_0}^{\mathrm{T}}\xi\|_2\big) \cdot \sup_{(\delta,\Delta)\in B_{\mathscr{A}}(r)} \sum_{k=1}^K \|\Delta_{\mathscr{A}}^k\|_2 \le \frac{4r}{n}\mathbb{E}\big(\|\mathbb{X}_{\mathscr{A}_0}^{\mathrm{T}}\xi\|_2\big), \end{aligned} \tag{34}$$

where $\xi = (\xi_1, \ldots, \xi_n)^{\mathrm{T}}$ is a random vector of i.i.d. Rademacher variables that is independent of $\{\varepsilon_1, \ldots, \varepsilon_n\}$. By Jensen's and Khintchine inequalities [28], we have

$$\begin{aligned} \mathbb{E}\big(\|\mathbb{X}_{\mathscr{A}_0}^{\mathrm{T}}\xi\|_2\big) &\le \big[\mathbb{E}\big(\xi^{\mathrm{T}}\mathbb{X}_{\mathscr{A}_0}\mathbb{X}_{\mathscr{A}_0}^{\mathrm{T}}\xi\big)\big]^{1/2} = \bigg[\sum_{j\in\mathscr{A}_0} \mathbb{E}\bigg(\sum_{i=1}^n \xi_i x_{ij}\bigg)^2\bigg]^{1/2} \\ &\le \bigg(\sum_{j\in\mathscr{A}_0}\sum_{i=1}^n x_{ij}^2\bigg)^{1/2} = \bigg(\sum_{i=1}^n \sum_{j\in\mathscr{A}_0} x_{ij}^2\bigg)^{1/2} \le \sqrt{n(s+1)M_{\mathscr{A}}}. \end{aligned}$$

It follows from (34) that $\mathbb{E}[z(r)] \le 4r\sqrt{(s+1)M_{\mathscr{A}}/n}$. The first part of the lemma then follows from (33).

Let $F(\delta,\Delta) = Q_n(\alpha^* + \delta, \beta^* + \Delta) - Q_n(\alpha^*, \beta^*)$. To prove the second inequality of the lemma, it suffices to note that

$$\inf_{(\delta,\Delta)\in S_{\mathscr{A}}(r)} F(\delta,\Delta) \ge \inf_{(\delta,\Delta)\in S_{\mathscr{A}}(r)} \mathbb{E}\big[Q_n(\alpha^* + \delta, \beta^* + \Delta) - Q_n(\alpha^*, \beta^*)\big] - z(r),$$

and that by (28) and the mean value theorem, we have for some $\bar{u}_{ik,t}$ between 0 and t such that

$$\begin{aligned} &\inf_{(\delta,\Delta)\in S_{\mathscr{A}}(r)} \mathbb{E}\big[Q_n(\alpha^* + \delta, \beta^* + \Delta) - Q_n(\alpha^*, \beta^*)\big] \\ &= \inf_{(\delta,\Delta)\in S_{\mathscr{A}}(r)} \frac{1}{nK} \sum_{i=1}^n \sum_{k=1}^K \int_0^{\delta_k + \mathbf{x}_i^{\mathrm{T}}\Delta} \big[F(\alpha_k^* + t) - F(\alpha_k^*)\big]\mathrm{d}t \\ &= \inf_{(\delta,\Delta)\in S_{\mathscr{A}}(r)} \frac{1}{nK} \sum_{i=1}^n \sum_{k=1}^K \int_0^{\delta_k + \mathbf{x}_i^{\mathrm{T}}\Delta} \big[t f(\alpha_k^* + \bar{u}_{ik,t})\big]\mathrm{d}t. \end{aligned} \tag{35}$$

Now for any $1 \le i \le n, 1 \le k \le K$ and $(\delta,\Delta) \in S_{\mathscr{A}}(r)$, we have

$$|\delta_k + \mathbf{x}_i^{\mathrm{T}}\Delta| \le \sqrt{1 + \|\mathbf{x}_{i\mathscr{A}}\|_2^2} \cdot \sqrt{\delta_k^2 + \|\Delta_{\mathscr{A}}\|_2^2} \le r\sqrt{(s+1)M_{\mathscr{A}}} \le \mathscr{U}_0.$$

It then follows from condition (C0) and (35) that

$$\begin{aligned}&\inf_{(\delta,\Delta)\in S_{\mathscr{A}}(r)} \mathbb{E}\big[Q_n(\alpha^*+\delta,\beta^*+\Delta)-Q_n(\alpha^*,\beta^*)\big]\\ &\geq \inf_{(\delta,\Delta)\in S_{\mathscr{A}}(r)} \frac{\underline{f}}{2nK}\sum_{i=1}^{n}\sum_{k=1}^{K}(\delta_k+\mathbf{x}_i^{\mathrm{T}}\Delta)^2 \geq \frac{\underline{f}}{2K}\underline{\mu} r^2.\end{aligned}$$

This completes the proof. □

Proof of Lemma 7. Note that Q_n is convex, but not differentiable. Denote the subdifferential of $Q_n(\alpha,\beta)$ at $(\widehat{\alpha}^{\mathrm{o}},\widehat{\beta}^{\mathrm{o}})$ by

$$\begin{aligned}\partial Q_n(\widehat{\alpha}^{\mathrm{o}},\widehat{\beta}^{\mathrm{o}}) = \Bigg\{(\zeta,\xi)\colon\ &\zeta_k = \frac{1-2\tau_k}{2K} - \frac{1}{2nK}\sum_{i=1}^{n}\mathrm{Sgn}(\hat{r}_{ik}),\ 1\leq k\leq K,\\ &\xi_j = \frac{1}{2n}\sum_{i=1}^{n}x_{ij}\Big(1-\frac{2}{K}\sum_{l=1}^{K}\tau_l\Big) - \frac{1}{2nK}\sum_{i=1}^{n}\sum_{l=1}^{K}\mathrm{Sgn}(\hat{r}_{il})x_{ij},\ 1\leq j\leq p\Bigg\}.\end{aligned}$$

By convexity of Q_n, for any $(\zeta,\xi)\in\partial Q_n(\widehat{\alpha}^{\mathrm{o}},\widehat{\beta}^{\mathrm{o}})$ and (α,β), we have

$$Q_n(\alpha,\beta)-Q_n(\widehat{\alpha}^{\mathrm{o}},\widehat{\beta}^{\mathrm{o}}) \geq \zeta^{\mathrm{T}}(\alpha-\widehat{\alpha}^{\mathrm{o}})+\xi^{\mathrm{T}}(\beta-\widehat{\beta}^{\mathrm{o}}).$$

Now by optimality of $(\widehat{\alpha}^{\mathrm{o}},\widehat{\beta}^{\mathrm{o}})$, we can take $\zeta=\mathbf{0}$ and $\xi_{\mathscr{A}}=\mathbf{0}$. It follows that

$$Q_n(\alpha,\beta)\geq Q_n(\widehat{\alpha}^{\mathrm{o}},\widehat{\beta}^{\mathrm{o}})+\sum_{j\in\mathscr{A}^{\mathrm{c}}}\xi_j(\beta_j-\hat{\beta}_j^{\mathrm{o}}). \tag{36}$$

Under event $\mathscr{E}_1$, we have $\max_{j\in\mathscr{A}^{\mathrm{c}}}|\hat{\beta}_j^{(0)}|\leq a_0\lambda\leq a_2\lambda$. Moreover, by condition (C3), we have

$$\min_{j\in\mathscr{A}}|\hat{\beta}^{(0)}| \geq \min_{j\in\mathscr{A}}|\beta_j^*| - \max_{j\in\mathscr{A}}|\hat{\beta}_j^{(0)}-\beta_j^*| \geq (a+1-a_0)\lambda \geq a\lambda.$$

Thus, under event $\mathscr{E}_1$, it follows from properties (P3) and (P4) of $p_\lambda(\cdot)$ that

$$p'_\lambda(|\hat{\beta}_j^{(0)}|)\geq a_1\lambda,\forall j\in\mathscr{A}^{\mathrm{c}} \quad\text{and}\quad p'_\lambda(|\hat{\beta}_j^{(0)}|)=0,\forall j\in\mathscr{A}.$$

Similarly, under event $\mathscr{E}_3$ and by the fact that $\widehat{\beta}^{\mathrm{o}}_{\mathscr{A}^{\mathrm{c}}}=\mathbf{0}$, it can be shown that

$$p'_\lambda(|\hat{\beta}_j^{\mathrm{o}}|)=0,\forall j\in\mathscr{A} \quad\text{and}\quad p'_\lambda(|\hat{\beta}_j^{\mathrm{o}}|)\geq a_1\lambda,\forall j\in\mathscr{A}^{\mathrm{c}}.$$

To this end, it can be seen from step (2.a) of the LLA algorithm that

$$(\widehat{\alpha}^{(1)},\widehat{\beta}^{(1)}) = \arg\min_{\alpha,\beta} Q_n(\alpha,\beta)+\sum_{j\in\mathscr{A}^{\mathrm{c}}}p'_\lambda(|\hat{\beta}_j^{(0)}|)|\beta_j|.$$

Now under $\mathscr{E}_2=\{\|\xi_{\mathscr{A}^{\mathrm{c}}}\|_\infty<a_1\lambda\}$, it follows from (36) that for any (α,β),

$$\begin{aligned}&\Big[Q_n(\alpha,\beta)+\sum_{j\in\mathscr{A}^{\mathrm{c}}}p'_\lambda(|\hat{\beta}_j^{(0)}|)|\beta_j|\Big]-\Big[Q_n(\widehat{\alpha}^{\mathrm{o}},\widehat{\beta}^{\mathrm{o}})+\sum_{j\in\mathscr{A}^{\mathrm{c}}}p'_\lambda(|\hat{\beta}_j^{(0)}|)|\hat{\beta}_j^{\mathrm{o}}|\Big]\\ &\geq \sum_{j\in\mathscr{A}^{\mathrm{c}}}\xi_j(\beta_j-\hat{\beta}_j^{\mathrm{o}})+\sum_{j\in\mathscr{A}^{\mathrm{c}}}p'_\lambda(|\hat{\beta}_j|^{(0)})|\beta_j| \qquad (37)\\ &\geq \sum_{j\in\mathscr{A}^{\mathrm{c}}}\big[p'_\lambda(|\hat{\beta}_j^{(0)}|)-|\xi_j|\big]|\beta_j|\geq 0.\end{aligned}$$

The leftmost hand side of the above inequality is strictly positive unless $\beta_{\mathscr{A}^{\mathrm{c}}}=\mathbf{0}$. Note that condition (C0) implies the uniqueness of the oracle estimator (See Section B of this appendix). It can be then seen that $(\widehat{\alpha}^{(1)},\widehat{\beta}^{(1)})$ coincides

with the oracle estimator. Now given that $(\widehat{\alpha}^{(1)},\widehat{\beta}^{(1)})$ is the oracle estimator, we show that $(\widehat{\alpha}^{(2)},\widehat{\beta}^{(2)})$ yielded by the LLA algorithm will still be the oracle estimator. To see it, note that under event $\mathscr{E}_2$,

$$p'_\lambda(|\hat{\beta}_j^{(1)}|)=0, \forall j\in\mathscr{A} \quad \text{and} \quad p'_\lambda(|\hat{\beta}_j^{(1)}|)\geq a_1\lambda, \forall j\in\mathscr{A}^{\mathrm{c}}.$$

By the LLA iteration, we have

$$(\widehat{\alpha}^{(2)},\widehat{\beta}^{(2)})=\arg\min_{\alpha,\beta} Q_n(\alpha,\beta)+\sum_{j\in\mathscr{A}^{\mathrm{c}}} p'_\lambda(|\hat{\beta}_j^{(1)}|)|\beta_j|.$$

Thus, we can follow similar arguments from (37) to show that under event $\mathscr{E}_3$, $(\widehat{\alpha}^{(2)},\widehat{\beta}^{(2)})$ is still the oracle estimator. This proves the lemma. □

Note that the above proof is slightly different from the general result (Theorems 1 and 2) in [18] since we need to deal with the intercept terms additionally.

Proof of Lemma 8. Consider a minimal ψ-cover of $B_{\mathscr{A}}(r)$ and denote this covering net by $\{(\delta^\ell,\Delta^\ell), \ell=1,\ldots,N_\psi\}\subset B_{\mathscr{A}}(r)$. For $j\in\mathscr{A}^{\mathrm{c}}$, define

$$U_{ij}(\delta,\Delta)=\frac{1}{K}\sum_{k=1}^K\big[I(r_{ik}^*\leq\delta_k+\mathbf{x}_i^{\mathrm{T}}\Delta)-\tau_k\big]x_{ij}-\frac{1}{K}\sum_{k=1}^K\big[I(r_{ik}^*\leq 0)-\tau_k\big]x_{ij},$$

where $r_{ik}^*=y_i-\alpha_k^*-\mathbf{x}_i^{\mathrm{T}}\beta^*=\varepsilon_i-\alpha_k^*$, $1\leq i\leq n$, $1\leq k\leq K$. Then it can be seen that

$$\gamma_j(r)=\sup_{(\delta,\Delta)\in B_{\mathscr{A}}(r)}\left|\frac{1}{n}\sum_{i=1}^n\big[U_{ij}(\delta,\Delta)-\mathbb{E}U_{ij}(\delta,\Delta)\big]\right|.$$

For any $(\delta,\Delta)\in B_{\mathscr{A}}(r)$ and $j\in\mathscr{A}^{\mathrm{c}}$, note that

$$|U_{ij}(\delta,\Delta)|\leq\frac{1}{K}\sum_{k=1}^K|I(r_{ik}^*\leq\delta_k+\mathbf{x}_i^{\mathrm{T}}\Delta)-I(r_{ik}^*\leq 0)|\cdot|x_{ij}|\leq M_{\mathscr{A}^{\mathrm{c}}}.$$

Let $p_{ik}=\Pr\big(-(\delta_k+\mathbf{x}_i^{\mathrm{T}}\Delta)_-<r_{ik}^*\leq(\delta_k+\mathbf{x}_i^{\mathrm{T}}\Delta)_+\big)$, $1\leq i\leq n$, $1\leq k\leq K$. It follows from the mean value theorem and condition (C0) that

$$p_{ik}=F(\alpha_k^*+(\delta_k+\mathbf{x}_i^{\mathrm{T}}\Delta)_+)-F(\alpha_k^*-(\delta_k+\mathbf{x}_i^{\mathrm{T}}\Delta)_-)\leq\bar{f}|\delta_k+\mathbf{x}_i^{\mathrm{T}}\Delta|.$$

By Cauchy–Schwarz inequality and the mean value theorem, we have

$$\begin{aligned}\operatorname{var}\big[U_{ij}(\delta,\Delta)\big]&\leq\frac{x_{ij}^2}{K}\sum_{k=1}^K\operatorname{var}\big[I\big(-(\delta_k+\mathbf{x}_i^{\mathrm{T}}\Delta)_-<r_{ik}^*\leq(\delta_k+\mathbf{x}_i^{\mathrm{T}}\Delta)_+\big)\big]\\&=\frac{x_{ij}^2}{K}\sum_{k=1}^K p_{ik}(1-p_{ik})\leq\frac{\bar{f}x_{ij}^2}{K}\sum_{k=1}^K|\delta_k+\mathbf{x}_i^{\mathrm{T}}\Delta|\leq\frac{\bar{f}M_{\mathscr{A}^{\mathrm{c}}}^2}{K}\sum_{k=1}^K|\delta_k+\mathbf{x}_{i\mathscr{A}}^{\mathrm{T}}\Delta_{\mathscr{A}}|.\end{aligned}$$

Let $\Delta_{\mathscr{A}}^k=(\delta_k,\Delta_{\mathscr{A}}^{\mathrm{T}})^{\mathrm{T}}$, $1\leq k\leq K$. By Cauchy–Schwarz inequality again, we get

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n\operatorname{var}\big[U_{ij}(\delta,\Delta)\big]&\leq\frac{1}{nK}\bar{f}M_{\mathscr{A}^{\mathrm{c}}}^2\sum_{i=1}^n\sum_{k=1}^K|\delta_k+\mathbf{x}_{i\mathscr{A}}^{\mathrm{T}}\Delta_{\mathscr{A}}|\\&\leq\frac{\bar{f}M_{\mathscr{A}^{\mathrm{c}}}^2}{K}\sum_{k=1}^K\Big[\frac{1}{n}\big(\Delta_{\mathscr{A}}^k\big)^{\mathrm{T}}\mathbb{X}_{\mathscr{A}_0}^{\mathrm{T}}\mathbb{X}_{\mathscr{A}_0}\Delta_{\mathscr{A}}^k\Big]^{1/2}\leq\bar{f}M_{\mathscr{A}^{\mathrm{c}}}^2\overline{\mu}^{1/2}r.\end{aligned}$$

Now applying Bernstein inequality, we have for any $(\delta,\Delta)\in B_{\mathscr{A}}(r)$ and $t>0$,

$$\Pr\left(\left|\frac{1}{n}\sum_{i=1}^n\big[U_{ij}(\delta,\Delta)-\mathbb{E}U_{ij}(\delta,\Delta)\big]\right|>t\right)\leq 2\exp\left(-\frac{nt^2}{2\bar{f}M_{\mathscr{A}^{\mathrm{c}}}^2\overline{\mu}^{1/2}r+\frac{4}{3}M_{\mathscr{A}^{\mathrm{c}}}t}\right).$$

Now for $1 \le \ell \le N_\psi$, let $B_\ell(\psi) = \{(\delta,\Delta)\colon \|\delta-\delta^\ell\|_2^2 + \|\Delta-\Delta^\ell\|_2^2 \le \psi^2, \Delta_{\mathscr{A}^{\rm c}} = \mathbf{0}\}$ be the ball centered at $(\delta^\ell,\Delta^\ell) \in B_{\mathscr{A}}(r)$ with radius ψ. For any $1 \le i \le n$, $1 \le k \le K$ and $(\delta,\Delta) \in B_\ell(\psi)$, note that

$$|(\delta_k + \mathbf{x}_i^{\rm T}\Delta) - (\delta_k^\ell + \mathbf{x}_i^{\rm T}\Delta^\ell)| \le \big(1 + \|\mathbf{x}_{i\mathscr{A}}\|_2^2\big)^{1/2}\psi \le \big[(s+1)M_{\mathscr{A}}\big]^{1/2}\psi.$$

For $1 \le i \le n$ and $j \in \mathscr{A}^{\rm c}$, let

$$V_{ij}(\delta^\ell,\Delta^\ell) = \frac{|x_{ij}|}{K}\sum_{k=1}^K \big[I\big(r_{ik}^* \le \delta_k^\ell + \mathbf{x}_i^{\rm T}\Delta^\ell + ((s+1)M_{\mathscr{A}})^{1/2}\psi\big) - I(r_{ik}^* \le \delta_k^\ell + \mathbf{x}_i^{\rm T}\Delta^\ell)\big].$$

Since the indicator function $I(u \le t)$ is nondecreasing in t, we have

$$\begin{aligned}
&\sup_{(\delta,\Delta)\in B_\ell(\psi)}\left|\frac{1}{n}\sum_{i=1}^n\big(U_{ij}(\delta,\Delta) - U_{ij}(\delta^\ell,\Delta^\ell) - \mathbb{E}\big[U_{ij}(\delta,\Delta) - U_{ij}(\delta^\ell,\Delta^\ell)\big]\big)\right| \\
&\le \frac{1}{nK}\sum_{i=1}^n |x_{ij}| \sum_{k=1}^K \big[I\big(r_{ik}^* \le \delta_k^\ell + \mathbf{x}_i^{\rm T}\Delta^\ell + ((s+1)M_{\mathscr{A}})^{1/2}\psi\big) - I(r_{ik}^* \le \delta_k^\ell + \mathbf{x}_i^{\rm T}\Delta^\ell) \\
&\qquad\qquad - \Pr\big(r_{ik}^* \le \delta_k^\ell + \mathbf{x}_i^{\rm T}\Delta^\ell - ((s+1)M_{\mathscr{A}})^{1/2}\psi\big) + \Pr(r_{ik}^* \le \delta_k^\ell + \mathbf{x}_i^{\rm T}\Delta^\ell)\big] \\
&\coloneqq I_1 + \frac{1}{n}\sum_{i=1}^n\big[V_{ij}(\delta^\ell,\Delta^\ell) - \mathbb{E}V_{ij}(\delta^\ell,\Delta^\ell)\big],
\end{aligned} \tag{38}$$

where

$$\begin{aligned}
I_1 = \frac{1}{nK}\sum_{i=1}^n |x_{ij}| \sum_{k=1}^K \big[&\Pr\big(r_{ik}^* \le \delta_k^\ell + \mathbf{x}_i^{\rm T}\Delta^\ell + ((s+1)M_{\mathscr{A}})^{1/2}\psi\big) \\
&- \Pr\big(r_{ik}^* \le \delta_k^\ell + \mathbf{x}_i^{\rm T}\Delta^\ell - ((s+1)M_{\mathscr{A}})^{1/2}\psi\big)\big].
\end{aligned}$$

By the mean value theorem, we have

$$I_1 \le \frac{1}{nK}\sum_{i=1}^n |x_{ij}| \sum_{k=1}^K \cdot 2\big((s+1)M_{\mathscr{A}}\big)^{1/2}\bar{f}\psi \le 2\bar{f}M_{\mathscr{A}^{\rm c}}\big((s+1)M_{\mathscr{A}}\big)^{1/2}\psi.$$

Similarly, it can be shown that $|V_{ij}(\delta^\ell,\Delta^\ell)| \le M_{\mathscr{A}^{\rm c}}$ and

$$\frac{1}{n}\sum_{i=1}^n \operatorname{var}(V_{ij}(\delta^\ell,\Delta^\ell)) \le \bar{f}M_{\mathscr{A}^{\rm c}}^2\big((s+1)M_{\mathscr{A}}\big)^{1/2}\psi.$$

It then follows from (38) and Bernstein inequality that for $1 \le \ell \le N_\psi$,

$$\begin{aligned}
&\Pr\left(\sup_{(\delta,\Delta)\in B_\ell(\psi)}\left|\frac{1}{n}\sum_{i=1}^n\big(U_{ij}(\delta,\Delta) - U_{ij}(\delta^\ell,\Delta^\ell) - \mathbb{E}\big[U_{ij}(\delta,\Delta) - U_{ij}(\delta^\ell,\Delta^\ell)\big]\big)\right| > t\right) \\
&\qquad \le 2\exp\left(-\frac{nt_1^2}{2\bar{f}M_{\mathscr{A}^{\rm c}}^2\big((s+1)M_{\mathscr{A}}\big)^{1/2}\psi + \frac{4}{3}M_{\mathscr{A}^{\rm c}}t_1}\right),
\end{aligned}$$

where $t_1 = \big[t - 2\bar{f}M_{\mathscr{A}^{\rm c}}\big((s+1)M_{\mathscr{A}}\big)^{1/2}\psi\big]_+$. The lemma then follows by noting that

$$\begin{aligned}
\Pr(\gamma_j(r) > t) &= \Pr\left(\sup_{(\delta,\Delta)\in B_{\mathscr{A}}(r)}\left|\frac{1}{n}\sum_{i=1}^n\big[U_{ij}(\delta,\Delta) - \mathbb{E}U_{ij}(\delta,\Delta)\big]\right| > t\right) \\
&\le \sum_{\ell=1}^{N_\psi}\Pr\left(\sup_{(\delta,\Delta)\in B_\ell(\psi)}\left|\frac{1}{n}\sum_{i=1}^n\big(U_{ij}(\delta,\Delta) - U_{ij}(\delta^\ell,\Delta^\ell)\right.\right. \\
&\qquad\qquad \left.\left. - \mathbb{E}\big[U_{ij}(\delta,\Delta) - U_{ij}(\delta^\ell,\Delta^\ell)\big]\big)\right| > \frac{t}{2}\right) \\
&\quad + \sum_{\ell=1}^{N_\psi}\Pr\left(\left|\frac{1}{n}\sum_{i=1}^n\big[U_{ij}(\delta^\ell,\Delta^\ell) - \mathbb{E}U_{ij}(\delta^\ell,\Delta^\ell)\big]\right| > \frac{t}{2}\right).
\end{aligned}$$

This completes the proof. □

Proof of Lemma 9. Let $U_i(\delta,\Delta)=\frac{1}{K}\sum_{k=1}^K\{\rho_{\tau_k}(r_{ik}^*-\delta_k-\mathbf{x}_i^{\mathrm{T}}\Delta)-\rho_{\tau_k}(r_{ik}^*)\}$, where $r_{ik}^*=y_i-\alpha_k^*-\mathbf{x}_i^{\mathrm{T}}\beta^*=\varepsilon_i-\alpha_k^*, 1\le i\le n, 1\le k\le K$. By Massart's concentration inequality, we have

$$\Pr(z(A,r)\ge\mathbb{E}[z(A,r)]+t)\le\exp\Big(-\frac{n^2t^2}{8b_n^2(A,r)}\Big),$$

where $b_n^2(A,r)=\sup_{(\delta,\Delta)\in B_A(r)}\sum_{i=1}^n\mathrm{var}(U_i(\delta,\Delta))$. It follows from Lipschitz continuity of the check loss that

$$\begin{aligned}b_n^2(A,r)&\le\frac{4}{K}\sup_{(\delta,\Delta)\in B_A(r)}\sum_{k=1}^K\sum_{i=1}^n(\delta_k+\mathbf{x}_i^{\mathrm{T}}\Delta)^2\\&=\frac{4}{K}\sup_{(\delta,\Delta)\in B_A(r)}\sum_{k=1}^K(\delta_k,\Delta_A^{\mathrm{T}})(\mathbf{1}_n,\mathbf{X}_A)^{\mathrm{T}}(\mathbf{1}_n,\mathbf{X}_A)(\delta_k,\Delta_A^{\mathrm{T}})^{\mathrm{T}}\\&\le\frac{4n}{K}\sup_{(\delta,\Delta)\in B_A(r)}\sum_{k=1}^K\bar{\zeta}[\delta_k^2+\|\Delta_A\|^2]\le 4n\bar{\zeta}r^2.\end{aligned}$$

Moreover, by the symmetrization procedure and the contraction principle, we obtain

$$\begin{aligned}\mathbb{E}[z(A,r)]&\le\frac{4}{nK}\sum_{k=1}^K\mathbb{E}\Big[\sup_{(\delta,\Delta)\in B_A(r)}\Big|\sum_{i=1}^n\xi_i(\delta_k+\mathbf{x}_{iA}^{\mathrm{T}}\Delta_A)\Big|\Big]\\&\le\frac{4}{nK}\mathbb{E}(\|(\mathbf{1}_n,\mathbf{X}_A)^{\mathrm{T}}\xi\|_2)\cdot\sup_{(\delta,\Delta)\in B_A(r)}\sum_{k=1}^K\|(\delta_k,\Delta_A^{\mathrm{T}})^{\mathrm{T}}\|_2\\&\le\frac{4r}{n}\mathbb{E}(\|(\mathbf{1}_n,\mathbf{X}_A)^{\mathrm{T}}\xi\|_2),\end{aligned}$$

where $\xi=(\xi_1,\ldots,\xi_n)^{\mathrm{T}}$ is a random vector of i.i.d. Rademacher variables, independent of $\varepsilon_1,\ldots,\varepsilon_n$. By Jensen's and Khintchine inequalities, we have

$$\begin{aligned}&\mathbb{E}(\|(\mathbf{1}_n,\mathbf{X}_A)^{\mathrm{T}}\xi\|_2)\le\{\mathbb{E}[\xi^{\mathrm{T}}(\mathbf{1}_n,\mathbf{X}_A)(\mathbf{1}_n,\mathbf{X}_A)^{\mathrm{T}}\xi]\}^{1/2}\\&=\Big[\sum_{j\in\{0\}\cup A}\mathbb{E}\Big(\sum_{i=1}^n\xi_ix_{ij}\Big)^2\Big]^{1/2}\le\Big(\sum_{j\in\{0\}\cup A}\sum_{i=1}^nx_{ij}^2\Big)^{1/2}\\&=\Big(\sum_{i=1}^n\sum_{j\in\{0\}\cup A}x_{ij}^2\Big)^{1/2}\le M\sqrt{n(|A|+1)}.\end{aligned}$$

It follows that $\mathbb{E}[z(A,r)]\le 4Mr\sqrt{(|A|+1)/n}$. The lemma then follows. □

Proof of Theorem 4. As in the proof of Theorem 3, split all models under consideration, $\{\hat{A}_\lambda:\lambda\in\Xi_n\}$, into three groups: $\{\hat{A}_\lambda:\lambda\in\Xi_n^-\}$, $\{\hat{A}_\lambda:\lambda\in\Xi_n^0\}$, and $\{\hat{A}_\lambda:\lambda\in\Xi_n^+\}$, where $\Xi_n^-=\{\lambda\in\Xi_n:\mathscr{A}\not\subset\hat{A}_\lambda\}$, $\Xi_n^0=\{\lambda\in\Xi_n:\hat{A}_\lambda=\mathscr{A}\}$ and $\Xi_n^+=\{\lambda\in\Xi_n:\mathscr{A}\subset\hat{A}_\lambda,\hat{A}_\lambda\neq\mathscr{A}\}$.

Let $\widehat{Q}_n^\lambda=(nK)^{-1}\sum_{i=1}^n\sum_{k=1}^K\rho_{\tau_k}(y_i-\hat{\alpha}_k^\lambda-\mathbf{x}_i^{\mathrm{T}}\hat{\beta}^\lambda)$, where $(\widehat{\alpha}^\lambda,\widehat{\beta}^\lambda)$ is the two-step LLA estimator to the folded concave penalized CQR (6) with lasso initialization. Also, let $Q_n^*=(nK)^{-1}\sum_{i=1}^n\sum_{k=1}^K\rho_{\tau_k}(y_i-\alpha_k^*-\mathbf{x}_i^{\mathrm{T}}\beta^*)$. For any $A\subset\{1,2,\ldots,p\}$, let $(\widehat{\alpha}^A,\widehat{\beta}^A)$ be the estimator obtained by fitting the canonical CQR to model A, i.e.,

$$(\widehat{\alpha}^A,\widehat{\beta}^A)=\mathop{\mathrm{argmin}}_{\alpha,\beta:\beta_{A^{\mathrm{c}}}=\mathbf{0}}\frac{1}{nK}\sum_{i=1}^n\sum_{k=1}^K\rho_{\tau_k}(y_i-\alpha_k-\mathbf{x}_i^{\mathrm{T}}\beta).\tag{39}$$

Define $\widehat{Q}_n^A=(nK)^{-1}\sum_{i=1}^n\sum_{k=1}^K\rho_{\tau_k}(y_i-\hat{\alpha}_k^A-\mathbf{x}_i^{\mathrm{T}}\hat{\beta}^A)$. For any $\lambda\in\Xi_n$, recall that $\hat{A}_\lambda=\{1\le j\le p:\hat{\beta}_j^\lambda\neq 0\}$ corresponds to the active set of the two-step LLA estimator $(\widehat{\alpha}^\lambda,\widehat{\beta}^\lambda)$. By optimality of $(\widehat{\alpha}^{\hat{A}_\lambda},\widehat{\beta}^{\hat{A}_\lambda})$ in (39), we have $\widehat{Q}_n^{\hat{A}_\lambda}\le\widehat{Q}_n^\lambda$.

Let $\mathscr{G}_n^+ = \{A : A \supset \mathscr{A}, A \neq \mathscr{A}, |A| \leq J_n\}$. It can be seen that $\{\hat{A}_\lambda : \lambda \in \Xi_n^+\} \subset \mathscr{G}_n^+$. For $r > 0$ and $A \in \mathscr{G}_n^+$, let $B_A(r) = \{(\delta, \Delta) \in \mathbb{R}^K \times \mathbb{R}^p : \|\delta\|^2 + \|\Delta_A\|^2 \leq r^2, \Delta_{A^c} = \mathbf{0}\}$ and $S_A(r) = \{(\delta, \Delta) \in \mathbb{R}^K \times \mathbb{R}^p : \|\delta\|^2 + \|\Delta_A\|^2 = r^2, \Delta_{A^c} = \mathbf{0}\}$.

Case I: overfitted models. By Theorem 2, under the assumptions of this theorem, we have

$$P(\hat{A}_{\lambda_n} \neq \mathscr{A}) \to \mathscr{O}(1) \quad \text{as } n \to \infty.$$

Therefore, for any $\lambda \in \Xi_n^+$, we have

$$\begin{aligned}
&\Pr\left(\inf_{\lambda \in \Xi_n^+} [\mathrm{BIC}^{\mathrm{HL}}(\lambda) - \mathrm{BIC}^{\mathrm{HL}}(\lambda_{\mathrm{n}})] > 0 \right) \\
&= \Pr\left(\inf_{\lambda \in \Xi_n^+} \left[\log(\widehat{Q}_n^\lambda / \widehat{Q}_n^{\mathscr{A}}) + (|\hat{A}_\lambda| - s) \frac{C_n \log(p)}{n} \right] > 0 \right) + \mathscr{O}(1) \\
&\geq \Pr\left(\inf_{\lambda \in \Xi_n^+} \left[\log(\widehat{Q}_n^{\hat{A}_\lambda} / \widehat{Q}_n^{\mathscr{A}}) + (|\hat{A}_\lambda| - s) \frac{C_n \log(p)}{n} \right] > 0 \right) + \mathscr{O}(1),
\end{aligned}$$

where the last inequality follows from the fact that $\widehat{Q}_n^\lambda \geq \widehat{Q}_n^{\hat{A}_\lambda}$. Moreover, note that $\widehat{Q}_n^{\hat{A}_\lambda} \leq \widehat{Q}_n^{\mathscr{A}} \leq Q_n^*$ due to inclusion $\mathscr{A} \subset \hat{A}_\lambda$. We then apply the inequality $\log(1+x) \leq x, \forall x \geq 0$ to get

$$\log\left(\frac{\widehat{Q}_n^{\hat{A}_\lambda}}{\widehat{Q}_n^{\mathscr{A}}} \right) = -\log\left(\frac{\widehat{Q}_n^{\mathscr{A}}}{\widehat{Q}_n^{\hat{A}_\lambda}} \right) = -\log\left(1 + \frac{\widehat{Q}_n^{\mathscr{A}} - \widehat{Q}_n^{\hat{A}_\lambda}}{\widehat{Q}_n^{\hat{A}_\lambda}} \right) \geq -\frac{\widehat{Q}_n^{\mathscr{A}} - \widehat{Q}_n^{\hat{A}_\lambda}}{\widehat{Q}_n^{\hat{A}_\lambda}}.$$

Let $F(\delta, \Delta) = Q_n(\alpha^* + \delta, \beta^* + \Delta) - Q_n(\alpha^*, \beta^*)$ for $(\delta, \Delta) \in \mathbb{R}^K \times \mathbb{R}^p$. For each model $A \in \mathscr{G}_n^+$, let $r_A^* = 16K(M + \bar{\zeta}^{1/2})(\underline{f}\underline{\zeta})^{-1} \sqrt{(|A|+1)\log(p)/n}$. If we can show that $\inf_{(\delta,\Delta) \in S_A(r_A^*)} F(\delta, \Delta) > 0$, then by convexity of $\rho_\tau(\cdot)$, we must have $\|\hat{\alpha}^A - \alpha^*\|_2^2 + \|\hat{\beta}^A - \beta^*\|_2^2 \leq (r_A^*)^2$. Indeed, by Knight's identity (see (28) of the appendix) and the mean value theorem, we have

$$\begin{aligned}
&\inf_{(\delta,\Delta) \in S_A(r_A^*)} \mathbb{E}[F(\delta, \Delta)] \\
&= \inf_{(\delta,\Delta) \in S_A(r_A^*)} \frac{1}{nK} \sum_{i=1}^n \sum_{k=1}^K \int_0^{\delta_k + \mathbf{x}_i^{\mathrm{T}} \Delta} [F(\alpha_k^* + t) - F(\alpha_k^*)] \mathrm{d}t \\
&= \inf_{(\delta,\Delta) \in S_A(r_A^*)} \frac{1}{nK} \sum_{i=1}^n \sum_{k=1}^K \int_0^{\delta_k + \mathbf{x}_i^{\mathrm{T}} \Delta} [t f(\alpha_k^* + \bar{u}_{ik,t})] \mathrm{d}t.
\end{aligned} \tag{40}$$

Note that J_n satisfies $J_n^2 \log(p)/n = \mathscr{O}(1)$. Therefore, for any $1 \leq i \leq n, 1 \leq k \leq K$ and $(\delta, \Delta) \in S_A(r_A^*)$, we have

$$\begin{aligned}
&|\delta_k + \mathbf{x}_i^{\mathrm{T}} \Delta| \leq \sqrt{1 + \|\mathbf{x}_{iA}\|^2} \cdot \sqrt{\delta_k^2 + \|\Delta_A\|_2^2} \\
&\leq M r_A^* \sqrt{|A| + 1} = 16KM(M + \bar{\zeta}^{1/2})(\underline{f}\underline{\zeta})^{-1}(|A|+1)\sqrt{\log(p)/n} \\
&\leq 16KM(M + \bar{\zeta}^{1/2})(\underline{f}\underline{\zeta})^{-1}(J_n + 1)\sqrt{\log(p)/n} = \mathscr{O}(\mathscr{U}_0).
\end{aligned}$$

It follows from condition (C0) and (40) that

$$\inf_{(\delta,\Delta) \in S_A(r_A^*)} \mathbb{E}[F(\delta, \Delta)] \geq \inf_{(\delta,\Delta) \in S_A(r_A^*)} \frac{\underline{f}}{2nK} \sum_{k=1}^K \sum_{i=1}^n (\delta_k + \mathbf{x}_i^{\mathrm{T}} \Delta)^2 \geq \frac{\underline{f}}{2K} \underline{\zeta} (r_A^*)^2.$$

Therefore, by Lemma 9, with probability at least $1 - \exp\{-nt^2/[32\bar{\zeta}(r_A^*)^2]\}$, we have

$$\begin{aligned}
\inf_{(\delta,\Delta) \in S_A(r_A^*)} F(\delta, \Delta) &\geq \inf_{(\delta,\Delta) \in S_A(r_A^*)} \mathbb{E}[F(\delta, \Delta)] - z(A, r_A^*) \\
&\geq \frac{\underline{f}}{2K} \underline{\zeta} (r_A^*)^2 - 4M r_A^* \sqrt{\frac{|A|+1}{n}} - t.
\end{aligned}$$

Now take $t = 8r_A^* \sqrt{\bar{\zeta}(|A|+1)\log(p)/n}$. It follows that for each $A \in \mathscr{G}_n^+$, with probability at least $b_n^A = 1 - \exp\{-2(|A|+1)\log(p)\}$, we have

$$\inf_{(\delta,\Delta)\in S_A(r_A^*)} F(\delta,\Delta) \geq \frac{\underline{f}\underline{\zeta}(r_A^*)^2}{2K} - r_A^* \sqrt{\frac{|A|+1}{n}} \Big(8\sqrt{\bar{\zeta}\log(p)} + 4M \Big) > 0,$$

which immediately implies that

$$\|\hat{\alpha}^A - \alpha\|_2^2 + \|\hat{\beta}^A - \beta^*\|_2^2 \leq (r_A^*)^2.$$

Now by the Bonferroni inequality, we have $\|\hat{\alpha}^A - \alpha\|_2^2 + \|\hat{\beta}^A - \beta^*\|_2^2 \leq (r_A^*)^2$ for all $A \in \mathscr{G}_n^+$ simultaneously with probability at least

$$\begin{aligned} b_n &= 1 - \sum_{|A|=s+1}^{J_n} \binom{p-s}{|A|-s} (1 - b_n^A) \\ &= 1 - \sum_{k=1}^{J_n - s} \binom{p-s}{k} \exp\{-2(k+s+1)\log(p)\} \\ &\geq 1 - p^{-2(s+1)} \sum_{k=1}^{p-s} \binom{p-s}{k} \Big(\frac{1}{p^2}\Big)^k \\ &= 1 - p^{-2(s+1)} \Big[\Big(1 + \frac{1}{p^2}\Big)^{p-s} - 1 \Big] \to 1 \quad \text{as } p \to \infty. \end{aligned} \tag{41}$$

Now we derive the upper bound for $\sup_{A\in\mathscr{G}_n^+} |\widehat{Q}_n^A - Q_n^*|$. Let $\hat{\delta}^A = \hat{\alpha}^A - \alpha^*$ and $\hat{\Delta}^A = \hat{\beta}^A - \beta^*$. Observe that

$$\begin{aligned} |\widehat{Q}_n^A - Q_n^*| &= \Big| \frac{1}{nK} \sum_{i=1}^n \sum_{k=1}^K \{ \rho_{\tau_k}(y_i - \hat{\alpha}_k^A - \mathbf{x}_i^{\mathrm{T}} \hat{\beta}^A) - \rho_{\tau_k}(y_i - \alpha_k^* - \mathbf{x}_i^{\mathrm{T}} \beta^*) \} \Big| \\ &\leq \big| \mathbb{E}\{ F(\hat{\delta}^A, \hat{\beta}^A) \} \big| + z(A, r_A^*). \end{aligned}$$

Similarly, we have

$$\begin{aligned} \big| \mathbb{E}\{ F(\hat{\delta}^A, \hat{\beta}^A) \} \big| &= \Big| \frac{1}{nK} \sum_{i=1}^n \sum_{k=1}^K \int_0^{\hat{\delta}_k^A + \mathbf{x}_i^{\mathrm{T}} \hat{\Delta}^A} [F(\alpha_k^* + t) - F(\alpha_k^*)] \mathrm{d}t \Big| \\ &\leq \frac{1}{nK} \sum_{i=1}^n \sum_{k=1}^K \int_0^{|\hat{\delta}_k^A + \mathbf{x}_i^{\mathrm{T}} \hat{\Delta}^A|} [t f(\alpha_k^* + \bar{u}_{ik,t})] \mathrm{d}t \\ &\leq \frac{\bar{f}}{2nK} \sum_{i=1}^n \sum_{k=1}^K (\hat{\delta}^A + \mathbf{x}_i^{\mathrm{T}} \hat{\Delta}^A)^2 \\ &\leq \frac{1}{2} \bar{f}\bar{\zeta} (\|\hat{\delta}^A\|_2^2 + \|\hat{\Delta}^A\|_2^2) \leq \frac{1}{2} \bar{f}\bar{\zeta} (r_A^*)^2. \end{aligned}$$

It follows that with probability at least b_n,

$$\begin{aligned} |\widehat{Q}_n^A - Q_n^*| &\leq \frac{1}{2} \bar{f}\bar{\zeta} (r_A^*)^2 + 4M r_A^* \sqrt{\frac{|A|+1}{n}} + 8 r_A^* \sqrt{\frac{\bar{\zeta}(|A|+1)\log(p)}{n}} \\ &\leq \frac{128 K^2 (M + \bar{\zeta}^{1/2})^2}{\underline{f}\underline{\zeta}} \frac{(|A|+1)\log(p)}{n} \end{aligned}$$

holds for all $A \in \mathscr{G}_n^+$. Now going back to BIC, we have

$$\begin{aligned} &\Pr\Big(\inf_{\lambda \in \Xi_n^+} \Big[\log(\widehat{Q}_n^{\hat{A}_\lambda} / \widehat{Q}_n^{\mathscr{A}}) + (|\hat{A}_\lambda| - s) \frac{C_n \log(p)}{n} \Big] > 0 \Big) \\ &\geq \Pr\Big(\frac{C_n \log(p)}{n} - \sup_{A\in\mathscr{G}_n^+} \frac{\widehat{Q}_n^{\mathscr{A}} - \widehat{Q}_n^A}{(|A|-s)\widehat{Q}_n^A} > 0 \Big). \end{aligned}$$

Since $\mathbb{E}(|\varepsilon|)<\infty$, it follows that $\mathbb{E}(Q_n^*)\le \mathbb{E}(|\varepsilon|)+\sum_{k=1}^K|\alpha_k^*|/K<\infty$. Thus, we have

$$\widehat{Q}_n^A = Q_n^* - (Q_n^* - \widehat{Q}_n^A) = \mathscr{O}_P(1)$$

by noting that $J_n\log(p)=\mathscr{o}(n)$. Therefore, with probability at least b_n, we have

$$\sup_{A\in\mathscr{G}_n^+}\frac{\widehat{Q}_n^{\mathscr{A}}-\widehat{Q}_n^A}{(|A|-s)\widehat{Q}_n^A}\le \sup_{A\in\mathscr{G}_n^+}\frac{\widehat{Q}_n^*-\widehat{Q}_n^A}{(|A|-s)\widehat{Q}_n^A}=\mathscr{O}_P\Big(\frac{s\log(p)}{n}\Big).$$

Since $s=\mathscr{O}(1)$ and C_n diverges with n, we have $s\log(p)/n=\mathscr{o}(C_n\log(p)/n)$. It follows that

$$\Pr\left(\frac{C_n\log(p)}{n}-\sup_{A\in\mathscr{G}_n^+}\frac{\widehat{Q}_n^{\mathscr{A}}-\widehat{Q}_n^A}{(|A|-s)\widehat{Q}_n^A}>0\right)\to 1 \quad \text{as } n\to\infty,$$

which implies that $\Pr\big(\inf_{\lambda\in\Xi_n^+}[\mathrm{BIC}^{\mathrm{HL}}(\lambda)-\mathrm{BIC}^{\mathrm{HL}}(\lambda_{\mathrm{n}})]>0\big)\to 1$ as $n\to\infty$.

Case II: underfitted models. For any $\lambda\in\Xi_n^-$, similar to Case I, we have

$$\begin{aligned}&\Pr\Big(\inf_{\lambda\in\Xi_n^-}[\mathrm{BIC}^{\mathrm{HL}}(\lambda)-\mathrm{BIC}^{\mathrm{HL}}(\lambda_{\mathrm{n}})]>0\Big)\\ &\ge \Pr\left(\inf_{\lambda\in\Xi_n^-}\Big[\log(\widehat{Q}_n^{\hat{A}_\lambda}/\widehat{Q}_n^{\mathscr{A}})+(|\hat{A}_\lambda|-s)\frac{C_n\log(p)}{n}\Big]>0\right)+\mathscr{o}(1).\end{aligned}$$

Define $\mathrm{BIC}^{\mathrm{HL}}(A)=\log(n\widehat{Q}_n^A)+|A|C_n\log(p)/n$ and let $\mathscr{G}_n^-=\{A:|A|\le J_n,\mathscr{A}\not\subset A\}$. We can see that $\{\hat{A}_\lambda:\lambda\in\Xi_n^-\}\subset\mathscr{G}_n^-$. It suffices to show $\inf_{A\in\mathscr{G}_n^-}\mathrm{BIC}^{\mathrm{HL}}(A)>\mathrm{BIC}^{\mathrm{HL}}(\mathscr{A})$ with probability tending to one as $n\to\infty$. For any $A\in\mathscr{G}_n^-$, let $\bar{A}=A\cup\mathscr{A}$. Let $\theta=\min_{j\in\mathscr{A}}|\beta_j^*|$. Since $A\not\supset\mathscr{A}$, we must have $\|\hat{\alpha}^A-\alpha^*\|_2^2+\|\hat{\beta}^A-\beta^*\|_2\ge\theta^2$. However, since $\bar{A}\supset\mathscr{A}$ and $|\bar{A}|\le 2J_n$, using Lemma 9, we can similarly show as in Case I that $\|\hat{\alpha}^{\bar{A}}-\alpha^*\|_2^2+\|\hat{\beta}^{\bar{A}}-\beta^*\|_2^2\le\theta^2$ with probability at least $b_n^{\bar{A}}=1-\exp\{-2(|\bar{A}|+1)\log(p)\}$ as long as

$$\theta>8K(\underline{f}\underline{\zeta})^{-1}(2\sqrt{\bar{\zeta}\log(p)}+M)\sqrt{\frac{2J_n+1}{n}},$$

which is implied by the assumption $\sqrt{J_n\log(p)/n}=\mathscr{o}(\theta)$. It then follows that $\|\hat{\alpha}^{\bar{A}}-\alpha^*\|_2^2+\|\hat{\beta}^{\bar{A}}-\beta^*\|_2^2\le\theta^2$ holds for all $A\in\mathscr{G}_n^-$ with probability at least $\tilde{b}_n\to 1$ as $n\to\infty$, where

$$\tilde{b}_n=1-\sum_{|\bar{A}=s+1}^{2J_n}\binom{p-s}{|\bar{A}|-s}(1-b_n^{\bar{A}})\ge 1-p^{-2(s+1)}\Big[\Big(1+\frac{1}{p^2}\Big)^{p-s}-1\Big].$$

Therefore, there exists $a\in[0,1]$, $\bar{\alpha}^{\bar{A}}=a\hat{\alpha}^A+(1-a)\hat{\alpha}^{\bar{A}}$ and $\bar{\beta}^{\bar{A}}=a\hat{\beta}^A+(1-a)\hat{\beta}^{\bar{A}}$ such that $\|\bar{\alpha}^{\bar{A}}-\alpha^*\|_2^2+\|\bar{\beta}^{\bar{A}}-\beta^*\|_2^2=\theta^2$. By convexity of ρ_τ and the fact that $\widehat{Q}_n^A\ge\widehat{Q}_n^{\bar{A}}$, we have $\bar{Q}_n^{\bar{A}}=(nK)^{-1}\sum_{i=1}^n\sum_{k=1}^K\rho_{\tau_k}(y_i-\bar{\alpha}_k^{\bar{A}}-\mathbf{x}_i^{\mathrm{T}}\bar{\beta}^{\bar{A}})\le\widehat{Q}_n^A$. Note that $\widehat{Q}_n^{\bar{A}}\le\widehat{Q}_n^{\mathscr{A}}\le Q_n^*$. It follows that $\widehat{Q}_n^A-\widehat{Q}_n^{\bar{A}}\ge\bar{Q}_n^{\bar{A}}-Q_n^*$. For ease of notation, let $\bar{\delta}^{\bar{A}}=\bar{\alpha}^{\bar{A}}-\alpha^*$ and $\bar{\Delta}^{\bar{A}}=\bar{\beta}^{\bar{A}}-\beta^*$. It can be seen that

$$\bar{Q}_n^{\bar{A}}-Q_n^*\ge\mathbb{E}[F(\bar{\delta}^{\bar{A}},\bar{\Delta}^{\bar{A}})]-z(\bar{A},\theta).$$

Following similar arguments from Case I and noting that the support of $\bar{\beta}^{\bar{A}}$ is a subset of $\bar{A}$, we can show that with probability at least $\tilde{b}_n$, for all $A\in\mathscr{G}_n^-$, we have

$$\bar{Q}_n^{\bar{A}}-Q_n^*\ge\frac{\underline{f}}{2K}\underline{\zeta}\theta^2-4\theta\sqrt{\frac{|\bar{A}|+1}{n}}(2\sqrt{\bar{\zeta}\log(p)}+M).$$

Now for all $A \in \mathscr{G}_n^-$, applying the inequality $\log(1+x) \geq \min\{\log(2), x/2\}, \forall x \geq 0$, we have

$$\begin{aligned}
&\mathrm{BIC}^{\mathrm{HL}}(A) - \mathrm{BIC}^{\mathrm{HL}}(\bar{A}) = \log\left(1 + \frac{\widehat{Q}_n^A - \widehat{Q}_n^{\bar{A}}}{\widehat{Q}_n^{\bar{A}}}\right) + (|A| - |\bar{A}|)\frac{C_n \log(p)}{n} \\
&\geq \min\left\{\log(2), \frac{\bar{Q}_n^{\bar{A}} - Q_n^*}{\widehat{Q}_n^{\bar{A}}}\right\} - \frac{C_n s \log(p)}{n}.
\end{aligned}$$

Since $\sqrt{C_n s \log(p)/n} = o(\theta)$ and $\widehat{Q}_n^{\bar{A}} = O_P(1)$, it can be seen that with probability tending to one, we have $\inf_{A \in \mathscr{G}_n^-} \mathrm{BIC}^{\mathrm{HL}}(A) - \mathrm{BIC}^{\mathrm{HL}}(\bar{A}) > 0$. Following similar arguments as in Case I, we can show $\mathrm{BIC}^{\mathrm{HL}}(\bar{A}) \geq \inf_{S \supset \mathscr{A}, |S| \leq 2J_n} \mathrm{BIC}^{\mathrm{HL}}(S) \geq \mathrm{BIC}^{\mathrm{HL}}(\mathscr{A})$ with probability tending to one. Case II then follows by noting that

$$\begin{aligned}
&\inf_{A \in \mathscr{G}_n^-} [\mathrm{BIC}^{\mathrm{HL}}(A) - \mathrm{BIC}^{\mathrm{HL}}(\mathscr{A})] \\
&= \inf_{A \in \mathscr{G}_n^-} [\mathrm{BIC}^{\mathrm{HL}}(A) - \mathrm{BIC}^{\mathrm{HL}}(\bar{A}) + \mathrm{BIC}^{\mathrm{HL}}(\bar{A}) - \mathrm{BIC}^{\mathrm{HL}}(\mathscr{A})] \\
&\geq \inf_{A \in \mathscr{G}_n^-} [\mathrm{BIC}^{\mathrm{HL}}(A) - \mathrm{BIC}^{\mathrm{HL}}(\bar{A})].
\end{aligned}$$

□

APPENDIX B
NUMERICAL PROPERTIES OF THE CQR ORACLE SOLUTION

Recall that the CQR oracle estimator is obtained through regression on the true set of variables

$$(\widehat{\alpha}^{\mathrm{o}}, \widehat{\beta}^{\mathrm{o}}) := \underset{(\alpha,\beta):\, \beta_{\mathscr{A}^{\mathrm{c}}} = \mathbf{0}}{\arg\min} \sum_{k=1}^{K} w_k \sum_{i=1}^{n} \rho_{\tau_k}(y_i - \alpha_k - \mathbf{x}_i^{\mathrm{T}}\beta).$$

For ease of exposition, we will restrict the scope of variables under consideration to those in $\mathscr{A}$. Specifically, let $\mathbf{a} = \alpha$, $\mathbf{b} = \beta_{\mathscr{A}} \in \mathbb{R}^s$ and $\mathbf{z}_i = \mathbf{x}_{i\mathscr{A}}$, $i = 1, \ldots, n$. The oracle solution can be equivalently obtained through the following minimization problem

$$(\widehat{\mathbf{a}}, \widehat{\mathbf{b}}) := \underset{\mathbf{a},\mathbf{b}}{\arg\min} \sum_{k=1}^{K} w_k \sum_{i=1}^{n} \rho_{\tau_k}(y_i - a_k - \mathbf{z}_i^{\mathrm{T}}\mathbf{b}).$$

Now let $\mathbf{u}_k = \left(\mathbf{y} - a_k \mathbf{1}_n - \mathbf{Z}\mathbf{b}\right)_+$ and $\mathbf{v}_k = \left(\mathbf{y} - a_k \mathbf{1}_n - \mathbf{Z}\mathbf{b}\right)_-$, $k = 1, \ldots, K$, where $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)^{\mathrm{T}}$ and the positive and negative parts are taken componentwisely. Also, let $\mathbf{u} = (\mathbf{u}_1^{\mathrm{T}}, \ldots, \mathbf{u}_K^{\mathrm{T}})^{\mathrm{T}}$ and $\mathbf{v} = (\mathbf{v}_1^{\mathrm{T}}, \ldots, \mathbf{v}_K^{\mathrm{T}})^{\mathrm{T}}$. Then the above regression problem can be cast into the following linear program of standard form

$$\begin{aligned}
\text{minimize} \quad & c^{\mathrm{T}} x \\
\text{subject to} \quad & Ax = b \\
& x \succeq 0,
\end{aligned}$$

where $b = \mathbf{1}_K \otimes \mathbf{y}$, ($\otimes$: Kronecker product) and

$$\begin{aligned}
x &= \left(\mathbf{a}_+^{\mathrm{T}}, \mathbf{a}_-^{\mathrm{T}}, \mathbf{b}_+^{\mathrm{T}}, \mathbf{b}_-^{\mathrm{T}}, \mathbf{u}^{\mathrm{T}}, \mathbf{v}^{\mathrm{T}}\right)^{\mathrm{T}}, \\
c &= \left(\mathbf{0}_K^{\mathrm{T}}, \mathbf{0}_K^{\mathrm{T}}, \mathbf{0}_p^{\mathrm{T}}, \mathbf{0}_p^{\mathrm{T}}, w_1\tau_1\mathbf{1}_n^{\mathrm{T}}, \ldots, w_K\tau_K\mathbf{1}_n^{\mathrm{T}}, w_1(1-\tau_1)\mathbf{1}_n^{\mathrm{T}}, \ldots, w_K(1-\tau_K)\mathbf{1}^{\mathrm{T}}\right)^{\mathrm{T}}, \\
A &= \begin{pmatrix}
\mathbf{1}_n & \cdots & \mathbf{0} & -\mathbf{1}_n & \cdots & \mathbf{0} & \mathbf{Z} & -\mathbf{Z} & \mathbf{I}_n & -\mathbf{I}_n \\
\vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\mathbf{0} & \cdots & \mathbf{1}_n & \mathbf{0} & \cdots & -\mathbf{1}_n & \mathbf{Z} & -\mathbf{Z} & \mathbf{I}_n & -\mathbf{I}_n
\end{pmatrix}_{(nK)\times(2K+2s+2n)}.
\end{aligned}$$

Without loss of generality, assume that $\mathbf{1}_n \notin \mathrm{Span}(\mathbf{Z})$, where $\mathrm{Span}(\mathbf{Z})$ denotes the column span of $\mathbf{Z}$. Write

$$D = \begin{pmatrix} \mathbf{1}_n & \cdots & \mathbf{0} & \mathbf{Z} \\ \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \cdots & \mathbf{1}_n & \mathbf{Z} \end{pmatrix}_{(nK)\times(K+s)}.$$

The rows of D will be denoted by $d_i^{\mathrm{T}},\ i = 1,\ldots,nK$. Let $\mathscr{H}$ be the collection of $(K+s)$-element subsets of $\{1,\ldots,nK\}$. For $h \in \mathscr{H}$, let $D(h)$ denote the submatrix of D with rows $\{d_i^{\mathrm{T}}, i \in h\}$ and $b(h)$ be the $(K+s)$-vector with coordinates $\{b_i, i \in h\}$. We also let $\bar{h} = \{1,\ldots,n\}\backslash h$ for $h \in \mathscr{H}$. Let $H = \{h \in \mathscr{H}\colon |D(h)| \neq 0\}$. By similar arguments as in Section 6.2 of [3], one can verify that the vertices of the polyhedron $\{x\colon Ax = b, x \succeq 0\}$ are given by

$$\begin{aligned} &(\mathbf{a}^{\mathrm{T}}, \mathbf{b}^{\mathrm{T}})^{\mathrm{T}} = \big[D(h)\big]^{-1} b(h) \\ &\mathbf{u}(h) = \mathbf{v}(h) = \mathbf{0} \\ &\mathbf{u}(\bar{h}) = \left[b(\bar{h}) - D(\bar{h}) \begin{pmatrix} \mathbf{a} \\ \mathbf{b} \end{pmatrix} \right]_+ \\ &\mathbf{v}(\bar{h}) = \left[b(\bar{h}) - D(\bar{h}) \begin{pmatrix} \mathbf{a} \\ \mathbf{b} \end{pmatrix} \right]_- \end{aligned}$$

for all $h \in H$. According to the simplex algorithm (see, e.g., [29], Chapter 3), the optimal solution to this linear program is among the above set of vertices. Recall that y has a density with respect to the Lebesgue measure. It can be seen that with probability one, there are at most $K(K+s)$ zero residuals for which $y_i - \hat{a}_k - \mathbf{z}_i^{\mathrm{T}}\widehat{\mathbf{b}} = 0$, $1 \le i \le n$, $1 \le k \le K$, given each optimal solution $(\widehat{\mathbf{a}}, \widehat{\mathbf{b}})$. Otherwise, suppose that there exist $h \in H$ and $i \in \bar{h}$ such that $\mathbf{u}(i) = \mathbf{v}(i) = 0$. Then since $D(h)$ is non-singular, it follows that

$$b_i = d_i^{\mathrm{T}} \begin{pmatrix} \mathbf{a} \\ \mathbf{b} \end{pmatrix} = d_i^{\mathrm{T}} \big[D(h)\big]^{-1} b(h),$$

which implies that b_i is a linear combination of $b(h)$. By the assumption that y has a density and the structure of b, this occurs with probability zero unless $b_i = b_j$ for some $j \in h$. However, there are at most $(K-1)$ such i's for each $j \in h$. This means with probability one, at each vertex, there are at most $K(K+s)$ indices i for which $\mathbf{u}(i) = \mathbf{v}(i) = 0$.

APPENDIX C
CONVERGENCE CRITERION FOR THE ADMM ALGORITHM

We adopt the following convergence criterion recommended by [30] for the ADMM algorithm (Algorithm 1) we propose to solve the weighted lasso penalized CQR problem:

$$\left\| \begin{pmatrix} \mathbb{X}_1 \\ -\mathbb{X}_2 \end{pmatrix} \boldsymbol{\varphi}^r + \begin{pmatrix} \mathrm{vec}(\mathbf{Z}^r) \\ \boldsymbol{\gamma}^r \end{pmatrix} - \begin{pmatrix} \mathbf{Y} \\ \mathbf{0} \end{pmatrix} \right\|_2 \leq \varepsilon_1 \sqrt{nK+p}$$
$$+ \varepsilon_2 \cdot \max\left\{ \left\| \begin{pmatrix} \mathbb{X}_1 \\ -\mathbb{X}_2 \end{pmatrix} \boldsymbol{\varphi}^r \right\|_2, \left\| \begin{pmatrix} \mathrm{vec}(\mathbf{Z}^r) \\ \boldsymbol{\gamma}^r \end{pmatrix} \right\|_2, \|\mathbf{Y}\|_2 \right\},$$
$$\sigma \|\mathbb{X}_1^{\mathrm{T}}\{\mathrm{vec}(\mathbf{Z}^r) - \mathrm{vec}(\mathbf{Z}^{r-1})\} - \mathbb{X}_2^{\mathrm{T}}(\boldsymbol{\gamma}^r - \boldsymbol{\gamma}^{r-1})\|_2 \leq \varepsilon_1 \sqrt{p+K}$$
$$+ \varepsilon_2 \cdot \|\mathbb{X}_1^{\mathrm{T}} \mathrm{vec}(\mathbf{U}^r) - \mathbb{X}_2^{\mathrm{T}} \mathbf{v}^r\|_2,$$

where ε_1 and ε_2 are the tolerances taking small positive values.

REFERENCES

[1] R. Koenker and G. Bassett, "Regression quantiles," *Econometrica: Journal of the Econometric Society*, vol. 46, no. 1, pp. 33–50, Jan. 1978.

[2] K. Knight, "Limiting distributions for L_1 regression estimators under general conditions," *The Annals of Statistics*, vol. 26, no. 2, pp. 755–770, 1998.

[3] R. Koenker, *Quantile Regression*. Cambridge, United Kingdom: Cambridge University Press, 2005.

[4] H. Zou and M. Yuan, "Composite quantile regression and the oracle model selection theory," *The Annals of Statistics*, vol. 36, no. 3, pp. 1108–1126, 2008.

[5] R. Koenker, "A note on L-estimates for linear models," *Statistics & probability letters*, vol. 2, no. 6, pp. 323–325, 1984.

[6] Z. Zhao and Z. Xiao, "Efficient regressions via optimally combining quantile information," *Econometric theory*, vol. 30, no. 6, pp. 1272–1314, 2014.

[7] H. Zou, "The adaptive lasso and its oracle properties," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1418–1429, 2006.

[8] A. Belloni and V. Chernozhukov, "ℓ_1-penalized quantile regression in high-dimensional sparse models," *The Annals of Statistics*, vol. 39, no. 1, pp. 82–130, 02 2011. [Online]. Available: http://dx.doi.org/10.1214/10-AOS827

[9] B. Kai, R. Li, and H. Zou, "Local composite quantile regression smoothing: an efficient and safe alternative to local polynomial regression," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 72, no. 1, pp. 49–69, 2010.

[10] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society, Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[11] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Association*, vol. 96, no. 456, pp. 1348–1360, 2001. [Online]. Available: http://pubs.amstat.org/doi/abs/10.1198/016214501753382273

[12] C.-H. Zhang, "Nearly unbiased variable selection under minimax concave penalty," *The Annals of Statistics*, vol. 38, no. 2, pp. 894–942, 2010.

[13] L. Wang, Y. Wu, and R. Li, "Quantile regression for analyzing heterogeneity in ultra-high dimension," *Journal of the American Statistical Association*, vol. 107, no. 497, pp. 214–222, 2012.

[14] E. Candes and T. Tao, "The Dantzig selector: Statistical estimation when p is much larger than n," *The Annals of Statistics*, vol. 35, no. 6, pp. 2313–2351, Dec. 2007.

[15] P. J. Bickel, Y. Ritov, and A. B. Tsybakov, "Simultaneous analysis of Lasso and Dantzig selector," *The Annals of Statistics*, vol. 37, no. 4, pp. 1705–1732, Aug. 2009.

[16] A. Belloni, V. Chernozhukov, and L. Wang, "Square-root lasso: pivotal recovery of sparse signals via conic programming," *Biometrika*, vol. 98, no. 4, pp. 791–806, 2011.

[17] J. Fan and J. Lv, "Nonconcave penalized likelihood with NP-dimensionality," *IEEE Transactions on Information Theory*, vol. 57, no. 8, pp. 5467–5484, Aug. 2011.

[18] J. Fan, L. Xue, and H. Zou, "Strong oracle optimality of folded concave penalized estimation," *The Annals of Statistics*, vol. 42, no. 3, pp. 819–849, 06 2014.

[19] J. Lv and Y. Fan, "A unified approach to model selection and sparse recovery using regularized least squares," *The Annals of Statistics*, vol. 37, no. 6A, pp. 3498–3528, Dec. 2009.

[20] H. Zou and R. Li, "One-step sparse estimates in nonconcave penalized likelihood models," *The Annals of Statistics*, vol. 36, no. 4, pp. 1509–1533, Aug. 2008.

[21] Y. Kim, J.-J. Jeon *et al.*, "Consistent model selection criteria for quadratically supported risks," *The Annals of Statistics*, vol. 44, no. 6, pp. 2467–2496, 2016.

[22] E. R. Lee, H. Noh, and B. U. Park, "Model selection via bayesian information criterion for quantile regression models," *Journal of the American Statistical Association*, vol. 109, no. 505, pp. 216–229, 2014.

[23] Y. Gu, J. Fan, L. Kong, S. Ma, and H. Zou, "ADMM for high-dimensional sparse penalized quantile regression," *Technometrics*, vol. 60, no. 3, pp. 319–331, 2018. [Online]. Available: https://doi.org/10.1080/00401706.2017.1345703

[24] D. Pollard, "Empirical processes: Theory and applications," *NSF-CBMS Regional Conference Series in Probability and Statistics*, vol. 2, pp. i–86, 1990. [Online]. Available: http://www.jstor.org/stable/4153175

[25] P. Bühlmann and S. van de Geer, *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer, 2011.

[26] A. van der Vaart and J. Wellner, *Weak convergence and empirical processes*. Springer, New York, 1996.

[27] M. Ledoux and M. Talagrand, *Probability in Banach Spaces, Isoperimetry and Processes*, 1st ed. Springer-Verlag Berlin Heidelberg, 1991.

[28] U. Haagerup, "The best constants in the Khintchine inequality," *Studia Mathematica*, vol. 3, no. 70, pp. 231–283, 1981.

[29] D. Bertsimas and J. N. Tsitsiklis, *Introduction to linear optimization*. Athena Scientific, 1997.

[30] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.