Structural Change in Sparsity*

Sokbae Lee[†], Yuan Liao[‡], Myung Hwan Seo[§] and Youngki Shin[¶]

18 November 2014

Abstract

In the high-dimensional sparse modeling literature, it has been crucially assumed that the sparsity structure of the model is homogeneous over the entire population. That is, the identities of important regressors are invariant across the population and across the individuals in the collected sample. In practice, however, the sparsity structure may not always be invariant in the population, due to heterogeneity across different sub-populations. We consider a general, possibly non-smooth M-estimation framework, allowing a possible structural change regarding the identities of important regressors in the population. Our penalized M-estimator not only selects covariates but also discriminates between a model with homogeneous sparsity and a model with a structural change in sparsity. As a result, it is not necessary to know or pretest whether the structural change is present, or where it occurs. We derive asymptotic bounds on the estimation loss of the penalized M-estimators, and achieve the oracle properties. We also show that when there is a structural change, the estimator of the threshold parameter is super-consistent. If the signal is relatively strong, the rates of convergence can be further improved and asymptotic distributional properties of the estimators including the threshold estimator can be established using an adaptive penalization. The proposed methods are then applied to quantile regression and logistic regression models and are illustrated via Monte Carlo experiments.

Keywords: structural change, change-point, variable selection, quantile regression, high-dimensional M-estimation, sparsity, LASSO, SCAD

*This work was supported in part by the European Research Council (ERC-2009-StG-240910-ROMETA) and by the Social Sciences and Humanities Research Council of Canada (SSHRCC).

[†]Department of Economics, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul, 151-742, Republic of Korea. Email:sokbae@gmail.com.

[‡]Department of Mathematics, University of Maryland, College Park, MD 20742, USA. Email: yuanliao@umd.edu.

[§]Department of Economics, London School of Economics, Houghton Street, London, WC2A 2AE, UK. Email: m.seo@lse.ac.uk.

[¶]Department of Economics, University of Western Ontario, 1151 Richmond Street N, London, ON N6A 5C2, Canada. Email: yshin29@uwo.ca.

1 Introduction

Sparsity is one of the most fundamental conditions in high-dimensional regression models which assumes that only a relatively small portion of the regressors are active in the model. It has been crucial to assume that the sparsity structure of the model is homogeneous over the entire population. That is, the identities of contributing regressors (such as genes, control variables, and environmental variables) are the same across the population and across the individuals in the collected sample. Under this condition, various methods, such as Lasso, Dantzig selector, and folded-concave penalizations, have been developed to identify the contributing variables and their effects on the response variable. The literature includes, for instance, Tibshirani (1996); Fan and Li (2001); Zou and Hastie (2005); Candes and Tao (2007); Negahban et al. (2012); Bickel et al. (2009); Meinshausen and Yu (2009); Zhang (2010); Belloni et al. (2011); Bühlmann and van de Geer (2011), among many others. Due to the invariance of the sparsity structure in the population, we may call such a standard sparsity condition *homogeneous sparsity*.

In practice, however, the sparsity structure may not always be invariant in the population, due to the heterogeneity across different sub-populations. For instance, when analyzing high-dimensional gene expression data for disease classifications, the identities of contributing genes may depend on the environmental or demographical variables, e.g., exposed temperature, age, weights or received treatments. In analyzing the effects of macroeconomic variables on the GDP growth rates, the contributing regressors may depend on the level of the base-GDP (Lee et al., 2014). Let Q be an observed environmental variable, which divides the population into two sub-populations $\{Q > \tau_0\}$ and $\{Q \leq \tau_0\}$ for some unknown threshold parameter τ_0. We consider a high-dimensional sparse model where the sparsity structure (e.g., identities and effects of important or contributing regressors) may differ between the two sub-populations, which allows a possible structural change of the statistical model. In particular, we allow no structural change as a special case, which corresponds to the usual sparse model. Our framework is expected to be extendable to allow for multi-changes with more than two different sparsity structures in the population.

To describe our estimation framework, let $Y \in \mathbb{R}$ be a response variable, $Q \in \mathbb{R}$ be an environmental variable that determines a possible structural change, and $X \in \mathbb{R}^p$ be a p-dimensional vector of covariates. Here, Q can be a component of X, and p is potentially much larger than the sample size n. Let $\{(Y_i, Q_i, X_i) : i = 1, \ldots, n\}$ denote independent and identically distributed (i.i.d.) copies of (Y, Q, X). We consider a general possibly non-smooth M-estimation framework that includes non-differentiable losses (such as quantile regression) and binary response models (e.g., logistic regression) as special cases. A statistical model with a possible structural change in the sparsity can be described as follows: the model involves β_0 and $\theta_0 = \beta_0 + \delta_0$ as the sparse structural parameters respectively for the sub-populations $\{Q \le \tau_0\}$ and $\{Q > \tau_0\}$, where τ_0 is an unknown threshold value that determines the "boundary" of the sub-populations. The model is associated with a known loss function $\rho(t_1, t_2) : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^+$, which is assumed to be convex and Liptschitz continuous with respect to t_2 for each t_1. The unknown parameters $(\beta_0, \delta_0, \tau_0)$ are defined as a minimizer of the expected loss (for simplicity, we assume that there is a unique minimizer):

$$(\beta_0, \delta_0, \tau_0) \equiv \operatorname{argmin}_{(\beta,\delta)\in\mathcal{A},\tau\in\mathcal{T}} \mathbb{E}\left[\rho(Y, X^T\beta + X^T\delta 1\{Q > \tau\})\right],$$

where $\mathcal{A} \times \mathcal{T}$ is the parameter space for $\alpha_0 \equiv (\beta_0^T, \delta_0^T)^T$ and τ_0.

For instance, in quantile regression models, for certain known $\gamma \in (0,1)$,

$$Y = X^T\beta_0 + X^T\delta_0 1\{Q > \tau_0\} + U, \quad P(U \le 0|X, Q) = \gamma, \tag{1.1}$$

$$\rho(t_1, t_2) = (t_1 - t_2)(\gamma - 1\{t_1 - t_2 \le 0\}).$$

Here $\rho(\cdot,\cdot)$ is the "check function" for quantile regressions. For a sparse vector $v \in \mathbb{R}^p$, we denote the active set of v as

$$J(v) \equiv \{j : v_j \neq 0\}.$$

Write $\theta_0 \equiv \beta_0 + \delta_0$. Let β_{0J} and θ_{0J} respectively denote the subvectors of nonzero components of β_0 and θ_0. Accordingly, let $X_{J(\beta_0)}$ and $X_{J(\theta_0)}$ denote the subvectors of X whose indices are in $J(\beta_0)$ and $J(\theta_0)$. Then model (1.1) corresponds to the quantile regression model with a structural-change regarding the identifies and effects of the contributing regressors:

$$Y = \begin{cases} X_{J(\beta_0)}^T\beta_{0J} + U, & Q \le \tau_0, \\ X_{J(\theta_0)}^T\theta_{0J} + U, & Q > \tau_0, \end{cases}$$

where the identities of $X_{J(\beta_0)}$, $X_{J(\theta_0)}$, the change point τ_0, and regression coefficients β_{0J}, θ_{0J} are all unknown.

We consider estimating regression coefficients (β_0, θ_0) as well as the threshold parameter τ_0 and selecting the contributing regressors in each sub-population based on ℓ_1-penalized M-estimators. One of the strengths of our proposed procedure is that it does not require to know or pretest whether $\delta_0 = 0$ or not, that is, whether the population's sparsity structure and regression effects are invariant or not. Neither do we need to know whether the threshold τ_0 is present in the model in order to establish oracle properties for the prediction risk and the estimation rates. As a result, the usual high-dimensional M-estimation without structural change is nested as a special case. Technically, we allow the loss function to be possibly non-smooth, with the quantile regression as a leading example, which broadens the scope of applications for penalized M-estimation. Moreover, the objective function is non-convex with respect to the threshold parameter τ_0, which is another technical challenge to handle.

Our paper is closely related to the statistical literature on models with unknown change points (e.g., Tong (1990); Chan (1993); Hansen (2000); Pons (2003); Kosorok and Song (2007); Seijo and Sen (2011a,b)). However, the model being considered is different both conceptually and technically, as it involves two high-dimensional parameters β_0 and δ_0 with a change of sparsity at an unknown threshold value τ_0. Moreover, recent related works on high-dimensional models are found in Enikeeva and Harchaoui (2013); Chan et al. (2014), Frick et al. (2014) and Cho and Fryzlewicz (2014), but they do not consider structural changes in the sparsity or possibly non-smooth general loss functions as we do in this paper. One exception is Lee et al. (2014), who studied a high-dimensional Gaussian mean regression with a change point in a deterministic design. However, as is clear from Belloni and Chernozhukov (2011), unlike the mean regressions, sparse quantile regression analyzes the effects of active regressors on different parts of the conditional distribution of a response variable, which provides a different angle of studying the regression effects. Dealing with non-smooth penalized loss functions with an unknown change point calls for a different technique. We also consider random designs and several oracle properties. Here the meaning of "oracle" is enriched compared with that of the homogeneous sparsity: in our problem, it is unknown to us whether the structural change is present or if it is present, where it occurs.

As we shall show, with possibly non-smooth and non-quadratic loss functions, the impact of not knowing the threshold value τ_0 leads to an additional term $(\log p)(\log n)$ in the asymptotic bounds, on top of those in the existing literature (e.g., Bickel et al. (2009); Belloni and Chernozhukov (2011)), and thus a slightly slower rate of convergence. But with a relatively stronger signal, using an adaptive double penalization, we can achieve the oracle rate. Furthermore, we establish another oracle property in that the estimation error in estimating τ_0 does not affect the asymptotic distribution of the estimate of α_0 and vice versa.

The remainder of the paper is organized as follows. Section 2 provides an informal description of our model and the estimation methodology. Section 3 establishes conditions under which the proposed estimator is consistent in terms of its excess risk and the estimated $\widehat{\tau}$. In addition, we derive the rate of convergence of the ℓ_1 estimation error for $\widehat{\alpha}$ and achieve the super-convergence rate for $\widehat{\tau}$ in the presence of sparsity-structural-change. The same rate of convergence for the excess risk as well as the ℓ_1 estimation error for $\widehat{\alpha}$ can be achieved even when there is no structural change. Section 4 achieves the variable selection consistency, again regardless of the existence of structural changes on the sparsity. We also establish conditions under which our estimators of α_0 and τ_0 have the oracle property. Section 5 verifies all the regularity conditions on the loss function for quantile and logistic regression models. Finally Section 6 gives the results of some simulations. Appendices A, B, and C that contain the proofs of all the theoretical results.

Notation. Throughout the paper, we use $|v|_q$ for the ℓ_q norm for a vector v with $q = 0, 1, 2$. For two sequences a_n and b_n, we write $a_n \ll b_n$ and equivalently $b_n \gg a_n$ if $a_n = o(b_n)$. Let $\lambda_{\min}(A)$ denote the minimum eigenvalue of a matrix A. We use w.p.a.1 to mean "with probability approaching one." The true parameter vectors β_0 and δ_0 except τ_0 are implicitly indexed by the sample size n, and we allow that the dimensions of $J(\beta_0)$, $J(\delta_0)$, and $J(\theta_0)$ can go to infinity as $n \to \infty$. For simplicity, we omit their dependence on n in our notation.

2 The Model and Estimators

In this section, we describe our model and the proposed estimation methodology.

2.1 Model

Recall that $\rho : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^+$ is a loss function under consideration, whose analytical form is clear in specific models, and that the true parameters are defined as the unique minimizer of the expected loss:

$$(\beta_0, \delta_0, \tau_0) \equiv \text{argmin}_{(\beta,\delta)\in\mathcal{A},\tau\in\mathcal{T}} \mathbb{E}\left[\rho(Y, X^T\beta + X^T\delta 1\{Q > \tau\})\right], \tag{2.1}$$

where $\mathcal{A}$ and $\mathcal{T}$ denote the parameter spaces for (β_0, δ_0) and τ_0. The equation (2.1) is usually satisfied by statistical models with a properly chosen loss function (We shall use quantile and logistic regressions as the main examples). Moreover, for each $(\beta, \delta) \in \mathcal{A}$ and $\tau \in \mathcal{T}$, define $2p \times 1$ vectors:

$$\alpha \equiv (\beta^T, \delta^T)^T, \quad X(\tau) \equiv (X^T, X^T 1\{Q > \tau\})^T.$$

Let $\alpha_0 \equiv (\beta_0^T, \delta_0^T)^T$. Then $X^T\beta + X^T\delta 1\{Q > \tau\} = X(\tau)^T\alpha$, and thus we can write (2.1) more compactly as:

$$(\alpha_0, \tau_0) = \text{argmin}_{\alpha\in\mathcal{A},\tau\in\mathcal{T}} \mathbb{E}\left[\rho(Y, X(\tau)^T\alpha)\right]. \tag{2.2}$$

Note that the loss $\rho(Y, X(\tau)^T\alpha)$ is not convex in τ.

2.2 The Lasso Estimator

Suppose we observe i.i.d. samples $\{Y_i, X_i, Q_i\}_{i\le n}$. Let $X_i(\tau)$ and $X_{ij}(\tau)$ denote the i-th realization of $X(\tau)$ and j-th element of $X_i(\tau)$, respectively, $i = 1, \ldots, n$ and $j = 1, \ldots, 2p$. Motivated from (2.2), we estimate the unknown parameters via an ℓ_1-penalized M-estimation:

$$(\widehat{\beta}, \widehat{\delta}, \widehat{\tau}) \equiv (\widehat{\alpha}, \widehat{\tau}) \equiv \text{argmin}_{\alpha\in\mathcal{A},\tau\in\mathcal{T}} S_n(\alpha, \tau), \tag{2.3}$$

where

$$\begin{aligned} S_n(\alpha,\tau) &\equiv \frac{1}{n}\sum_{i=1}^{n} \rho(Y_i, X_i(\tau)^T\alpha) + \lambda_n \sum_{j=1}^{2p} D_j(\tau)|\alpha_j| \\ &\equiv \frac{1}{n}\sum_{i=1}^{n} \rho(Y_i, X_i^T\beta + X_i^T\delta 1\{Q_i > \tau\}) + \lambda_n \sum_{j=1}^{p} d_j|\beta_j| + \lambda_n \sum_{j=1}^{p} d_j(\tau)|\delta_j|. \end{aligned}$$

Here λ_n is the tuning parameter, $D_j(\tau) \equiv (\frac{1}{n}\sum_{i=1}^n X_{ij}(\tau)^2)^{1/2}$, $j = 1, \ldots, 2p$, are the data-dependent weights adequately balancing the regressors, and $X_{ij}(\tau) \equiv (X_{ij}, X_{ij}1\{Q_i > \tau\})$. Note that the weight $d_j \equiv (\frac{1}{n}\sum_{i=1}^n X_{ij}^2)^{1/2}$ regarding $|\beta_j|$ does not depend on τ, while the weight $d_j(\tau) \equiv (\frac{1}{n}\sum_{i=1}^n X_{ij}^2 1\{Q_i > \tau\})^{1/2}$ with respect to $|\delta_j|$ does, which takes into account the effect of the threshold τ on the parameter change δ.

Remark 2.1. It is worth noting that alternatively, one might penalize β and $\beta + \delta$ instead of β, δ. We opt to penalize δ directly since this formulation makes it convenient to identify the set of regressors whose effects may have structural changes. Specifically, if a component δ_j is identified to be nonzero, it implies that there is a structural change on the jth regressor; otherwise there is no change on its regressor. As a result, our formulation includes the usual sparse modeling without sparsity-structural-change as a special case, by allowing $\delta_0 = 0$. Note that when $\delta_0 = 0$,

$$X^T\beta_0 + X^T\delta_0 1\{Q > \tau_0\} = X^T\beta_0,$$

hence τ_0 is non-identifiable. Since in practice, we do not know *ex ante* whether $\delta_0 = 0$ holds, we employ the same ℓ_1-penalized M-estimation as in (2.3), and (2.4) below, which still penalizes both β and δ. We shall show that in this case β_0 and δ_0 can still be consistently estimated, and their zero components can be identified.

2.3 The Estimator with the Oracle Property

When the signal strength is relatively strong, we can achieve the selection consistency. After the Lasso-step in (2.3), we employ an adaptively weighted ℓ_1-penalization, based on a local linear approximation (LLA) to the folded-concave penalty. Consider an objective function:

$$\widetilde{S}_n(\alpha) \equiv \frac{1}{n}\sum_{i=1}^{n}\rho(Y_i, X_i(\widehat{\tau})^T\alpha) + \mu_n \sum_{j=1}^{2p} w_j D_j(\widehat{\tau})|\alpha_j|, \tag{2.4}$$

where $\widehat{\tau}$ is the first-step estimator obtained from (2.3). The weights $\{w_j\}$ are determined through an LLA algorithm (Zou and Li (2008)) of the SCAD penalties. In usual sparse models without structural changes, algorithms of this type have been shown to achieve a fixed point that pertains the oracle properties (see Fan et al. (2014b)). In our context, we show that estimating τ_0 does not affect the oracle properties, no matter whether τ_0 is identifiable or not.

For some tuning parameter μ_n and for $j = 1, ..., p$, let w_j be the LLA of the SCAD-weight, namely,

$$w_j \equiv \begin{cases} 1, & |\widehat{\alpha}_j| < \mu_n \\ 0, & |\widehat{\alpha}_j| > a\mu_n \\ \frac{a\mu_n - |\widehat{\alpha}_j|}{\mu_n(a-1)} & \mu_n \leq |\widehat{\alpha}_j| \leq a\mu_n. \end{cases}$$

where $\widehat{\alpha}_j$ is the first-step estimator obtained from (2.3). Here $a > 1$ is some prescribed constant, and $a = 3.7$ is often used in the literature (e.g., Fan and Li (2001) and Loh and Wainwright (2013)).

We now define $\widetilde{\alpha}$ to be the global minimizer of $\widetilde{S}_n(\alpha)$ on the parameter space $\mathcal{A}$ of α:

$$\widetilde{\alpha} \equiv \arg\min_{\alpha\in\mathcal{A}} \widetilde{S}_n(\alpha). \tag{2.5}$$

The objective function is now convex, which facilities the computations. Once the asymptotically oracle estimator $\widetilde{\alpha}$ is obtained, we can improve upon the first Lasso estimator $\widehat{\tau}$. Define $\widetilde{\tau}$ to be

$$\widetilde{\tau} \equiv \underset{\tau\in\mathcal{T}}{\text{argmin}} \frac{1}{n}\sum_{i=1}^{n}\rho\left(Y_i, X_i(\tau)^T\widetilde{\alpha}\right). \tag{2.6}$$

We will establish that the estimators $\widetilde{\alpha}$ and $\widetilde{\tau}$ have the oracle properties. In the literature (see, e.g, Fan and Li (2001), Fan et al. (2014b)), an estimator is said to have the oracle property if it has the same asymptotic distribution as the infeasible oracle estimator. In our setup, an oracle knows $J(\beta_0)$, $J(\delta_0)$, τ_0 (if $\delta_0 \neq 0$), as well as whether $\delta_0 \neq 0$ or not. However, none of them are known to us. Hence, the meaning of the oracle property is enriched here compared with that of the homogeneous sparsity.

2.4 The Computation Algorithm

Numerically, for each fixed $\tau \in \mathcal{T}$, minimizing $S_n(\alpha, \tau)$ over $\alpha \in \mathcal{A}$ is a standard Lasso problem, and many efficient algorithms are available for various loss functions in the literature. Let $\widehat{\alpha}(\tau) = \text{argmin}_{\alpha \in \mathcal{A}} S_n(\alpha, \tau)$. Since $S_n(\widehat{\alpha}(\tau), \tau)$ takes on less than n distinct values, $\widehat{\tau}$ can be defined uniquely as

$$\widehat{\tau} = \arg\min_{\tau \in \widetilde{\mathcal{T}}_n} S_n(\widehat{\alpha}(\tau), \tau),$$

where $\widetilde{\mathcal{T}}_n \equiv \mathcal{T} \cap \{Q_1, ..., Q_n\}$. Hence the computation algorithm can be summarized as follows:

Step 1 For each $k = 1, ..., n$, set $\tau_k = Q_k$. For each $k = 1, ..., n$ such that $\tau_k \in \widetilde{\mathcal{T}}_n$, solve the Lasso problem:

$$\widehat{\alpha}(\tau_k) = \text{argmin}_{\alpha \in \mathcal{A}} S_n(\alpha, \tau_k).$$

Step 2 Set

$$k^* = \text{argmin}_{k=1,...,n: \tau_k \in \widetilde{\mathcal{T}}_n} S_n(\widehat{\alpha}(\tau_k), \tau_k), \quad \widehat{\alpha} = \widehat{\alpha}(\tau_{k^*}), \quad \widehat{\tau} = \tau_{k^*}$$

Step 3 Solve the LLA algorithm with $\widehat{\alpha}$ and $\widehat{\tau}$ obtained in step 2 to obtain $\widetilde{\alpha}$.

Step 4 Obtain $\widetilde{\tau}$ with $\widetilde{\alpha}$ obtained in step 3:

$$\widetilde{\tau} = \underset{\tau \in \widetilde{\mathcal{T}}}{\text{argmin}} \frac{1}{n} \sum_{i=1}^{n} \rho\left(Y_i, X_i(\tau)^T \widetilde{\alpha}\right).$$

In particular, steps 2 and 4 require only at most n function evaluations. If n is very large, $\widetilde{\mathcal{T}}_n$ can be approximated by a grid. For some $N < n$, let $Q_{(j)}$ denote the (j/N)th quantile of the sample $\{Q_1, ..., Q_n\}$, and let $\mathcal{T}_N = \mathcal{T} \cap \{Q_{(1)}, ..., Q_{(N)}\}$. Then in step 2, $\widehat{\tau}_N = \arg\min_{\tau \in \mathcal{T}_N} S_n(\widehat{\alpha}(\tau), \tau)$ is a good approximation to $\widehat{\tau}$ and the same applies to step 4.

3 Theoretical Properties of the Lasso Estimator

3.1 Assumptions

In this subsection, we collect regularity conditions that are needed to develop our theoretical results. Let X_{ij} denote the jth element of X_i and $\mathcal{T}_0 \subset \mathcal{T}$ a neighborhood of τ_0.

Assumption 3.1 (Setting). *(i) The data $\{(Y_i, X_i, Q_i)\}_{i=1}^n$ are independent and identically distributed with $\mathbb{E}|X_{ij}|^m \leq \frac{m!}{2} K_1^{m-2}$ for all j and some $K_1 < \infty$.*

(ii) $\alpha \in \mathcal{A} \equiv \{\alpha : |\alpha|_\infty \leq M_1\}$ for some $M_1 < \infty$, and $\tau \in \mathcal{T} \equiv [\underline{\tau}, \overline{\tau}]$, where the probability of $\{Q < \underline{\tau}\}$ and that of $\{Q > \overline{\tau}\}$ are strictly positive.

(iii) There exist universal constants $\underline{D} > 0$ and $\overline{D} > 0$ such that with probability approaching one,

$$0 < \underline{D} \leq \min_{j \leq 2p} \inf_{\tau \in \mathcal{T}} D_j(\tau) \leq \max_{j \leq 2p} \sup_{\tau \in \mathcal{T}} D_j(\tau) \leq \overline{D} < \infty.$$

(iv) There exists $\mathcal{T}_0$ such that $\sup_{j \leq p} \sup_{\tau \in \mathcal{T}_0} \mathbb{E}[X_j^2 | Q = \tau] < \infty$.

Condition (i) imposes mild moment restrictions on X. The compact parameter space in condition (ii) is standard in the literature on change-point and threshold models (e.g., Seijo and Sen (2011a,b)). Condition (iii) requires that each regressor be of the same magnitude uniformly over the threshold τ. As the data-dependent weights $D_j(\tau)$ are the sample second moments of the regressors, it is not stringent to assume them to be bounded away from both zero and infinity, given the well-behaved population counterparts. Condition (iv) assumes that the conditional expectation of $\mathbb{E}[X_j^2 | Q = \cdot]$ is bounded on $\mathcal{T}_0$ uniformly in j.

We paraphrase Assumption 3.2, which restricts the distribution of Q, before we state it. Condition (i) imposes a weak restriction on the distribution of Q, condition (ii) implies that $\mathbb{P}\{|Q - \tau_0| < \varepsilon\} > 0$ for any $\varepsilon > 0$, and condition (iii) requires that the conditional distribution of Q given X satisfy some weak restrictions.

Assumption 3.2 (Distribution of Q). *(i) $\mathbb{P}(\tau_1 < Q \leq \tau_2) \leq K_2(\tau_2 - \tau_1)$ for any $\tau_1 < \tau_2$ and some positive $K_2 < \infty$.*

(ii) Q has a density function that is continuous and bounded away from zero on $\mathcal{T}_0$.

(iii) The conditional distribution of Q given $\tilde{X}$ has a density function $f_{Q|\tilde{X}}(q|\tilde{x})$ that is bounded uniformly for $q \in \mathcal{T}_0$ and $\tilde{x}$, where $\tilde{X}$ denotes the all the components of X excluding Q in case that Q is an element of X.

We now state assumptions with respect to the objective function. Recall that $\theta_0 \equiv \beta_0 + \delta_0$, and let β, δ, and $\theta \equiv \beta + \delta$ denote the corresponding generic parameters. Also, recall that when $Q \leq \tau_0$, $X(\tau_0)^T \alpha_0 = X^T \beta_0$, while when $Q > \tau_0$, $X(\tau_0)^T \alpha_0 = X^T \theta_0$. Hence we define the "prediction balls" with radius r and corresponding centers as follows:

$$\begin{aligned} \mathcal{B}(\beta_0, r) &= \{\beta \in \mathcal{B} \subset \mathbb{R}^p : \mathbb{E}[(X^T(\beta - \beta_0))^2 1\{Q \leq \tau_0\}] \leq r^2\}, \\ \mathcal{G}(\theta_0, r) &= \{\theta \in \mathcal{G} \subset \mathbb{R}^p : \mathbb{E}[(X^T(\theta - \theta_0))^2 1\{Q > \tau_0\}] \leq r^2\}, \end{aligned} \tag{3.1}$$

where $\mathcal{B}$ and $\mathcal{G}$ are parameter spaces for β_0 and θ_0, respectively, which can be induced from $\mathcal{A}$. For a constant $\eta > 0$, define

$$r_1(\eta) \equiv \sup_r \Big\{ r : \mathbb{E}\left(\left[\rho\left(Y, X^T\beta\right) - \rho\left(Y, X^T\beta_0\right)\right] 1\left\{Q \leq \tau_0\right\}\right) \\ \geq \eta \mathbb{E}[(X^T(\beta - \beta_0))^2 1\{Q \leq \tau_0\}] \text{ for all } \beta \in \mathcal{B}(\beta_0, r) \Big\}$$

and

$$r_2(\eta) \equiv \sup_r \Big\{ r : \mathbb{E}\left(\left[\rho\left(Y, X^T\theta\right) - \rho\left(Y, X^T\theta_0\right)\right] 1\left\{Q > \tau_0\right\}\right) \\ \geq \eta \mathbb{E}[(X^T(\theta - \theta_0))^2 1\{Q > \tau_0\}] \text{ for all } \theta \in \mathcal{G}(\theta_0, r) \Big\}.$$

Note that $r_1(\eta)$ and $r_2(\eta)$ are the maximal radiuses over which the excess risk can be bounded below by the quadratic loss on $\{Q \leq \tau_0\}$ and $\{Q > \tau_0\}$, respectively.

Assumption 3.3 (Objective Function)**.** *(i) Let $\mathcal{Y}$ denote the support of Y. There is a Liptschitz constant $L > 0$ such that for all $y \in \mathcal{Y}$, $\rho(y, \cdot)$ is convex, and*

$$|\rho(y, t_1) - \rho(y, t_2)| \leq L|t_1 - t_2|, \forall t_1, t_2 \in \mathbb{R}.$$

(ii) For all $\alpha \in \mathcal{A}$, almost surely,

$$\mathbb{E}\left[\rho(Y, X(\tau_0)^T\alpha) - \rho(Y, X(\tau_0)^T\alpha_0)|Q\right] \geq 0,$$

(iii) There exist constants $\eta^* > 0$ and $r^* > 0$ such that $r_1(\eta^*) \geq r^*$ and $r_2(\eta^*) \geq r^*$.

(iv) There is a constant $c_0 > 0$ such that for all $\tau \in \mathcal{T}_0$,

$$\mathbb{E}\left[\left(\rho\left(Y, X^T\theta_0\right) - \rho\left(Y, X^T\beta_0\right)\right) 1\left\{\tau < Q \leq \tau_0\right\}\right] \geq c_0 \mathbb{E}\left[\left(X^T\left(\beta_0 - \theta_0\right)\right)^2 1\left\{\tau < Q \leq \tau_0\right\}\right], \\ \mathbb{E}\left[\left(\rho\left(Y, X^T\beta_0\right) - \rho\left(Y, X^T\theta_0\right)\right) 1\left\{\tau_0 < Q \leq \tau\right\}\right] \geq c_0 \mathbb{E}\left[\left(X^T\left(\beta_0 - \theta_0\right)\right)^2 1\left\{\tau_0 < Q \leq \tau\right\}\right].$$

In this paper, we focus on a convex Lipchitz loss function, which is assumed in condition (i). It is possible to relax the convexity and impose a "restricted strong convexity condition" as in Loh and Wainwright (2013). For simplicity, we focus on the case of a convex loss, which is satisfied by our leading examples. However, unlike the framework of M-estimation in Negahban et al. (2012) and Loh and Wainwright (2013), we do allow $\rho(t_1, t_2)$ to be non-differentiable, which admits quantile regression as a special case.

Condition (ii) is a weak condition given that

$$\mathbb{E}\left[\rho(Y, X(\tau)^T\alpha) - \rho(Y, X(\tau_0)^T\alpha_0)\right] \geq 0,$$

for any $\alpha \in \mathcal{A}$ and $\tau \in \mathcal{T}$. Condition (iii) requires that the excess risk can be bounded below by a quadratic function locally. Condition (iv) is in the same spirit as Condition (iii). Conditions (iii) and (iv), combined with the convexity of $\rho(Y, \cdot)$, helps us derive the rates of convergence (in the ℓ_1 norm) of the Lasso estimators of (α_0, τ_0). We shall provide primitive sufficient conditions for Assumption 3.3 for the quantile and logistic regression models in Section 5.

Remark 3.1. Condition (iii) of Assumption 3.3 is similar to *the restricted nonlinear impact (RNI)* condition of Belloni and Chernozhukov (2011). One may consider an alternative formulation as in van de Geer (2008) and Bühlmann and van de Geer (2011) (Chapter 6), which is known as the *margin condition*. But their margin condition needs to be adjusted to account for structural changes as in Condition (iv). It would be an interesting future research topic to develop a general theory of high-dimensional M-estimation with an unknown sparsity-structural-change with general margin conditions.

The following assumptions are needed to deal with the case when $\delta_0 \neq 0$.

Assumption 3.4 (Structural Change)**.** *Suppose that* $\delta_0 \neq 0$.

(i) $\mathbb{E}\left[\left(X^T\delta_0\right)^2 | Q = \tau\right] \leq M_2|\delta_0|_2^2$ *for all* $\tau \in \mathcal{T}$ *and for some* M_2 *satisfying* $0 < M_2 < \infty$.

(ii) For the same c_0 *in Assumption 3.3 (iv), we have that* $\mathbb{E}[(X^T\delta_0)^2|Q = \tau] \geq c_0$ *for all* $\tau \in \mathcal{T}_0$.

(iii) There exists $M_3 > 0$ *such that either* $M_3^{-1} \leq \mathbb{E}[(X^T\delta_0)^2|Q = \tau] \leq M_3$ *or* $M_3^{-1}|\delta_0|_2^2 \leq \mathbb{E}[(X^T\delta_0)^2|Q = \tau] \leq M_3|\delta_0|_2^2$ *holds for all* $\tau \in \mathcal{T}_0$.

Assumption 3.4 is concerned with $\mathbb{E}[\left(X^T\delta_0\right)^2 | Q = \tau]$, which is an important quantity to develop asymptotic results when $\delta_0 \neq 0$. Condition (i) puts some weak upper bound on $\mathbb{E}[\left(X^T\delta_0\right)^2 | Q = \tau]$ for all τ globally. Conditions (ii) and (iii) are local conditions with respect to τ. Condition (ii) is satisfied, for example, when Q has a density function that is bounded away from zero in a neighborhood of τ_0 and $q \mapsto \mathbb{E}[(X^T\delta_0)^2|Q = q]$ is uniformly continuous and strictly positive at $q = \tau_0$. Recall that the dimension of nonzero elements of δ_0 can grow with n. Condition (iii) requires that the growth rate of $\mathbb{E}[(X^T\delta_0)^2|Q = \tau]$ be uniform in τ, thereby implying that $\sup_{\tau\in\mathcal{T}_0} \mathbb{E}[(X^T\delta_0)^2|Q = \tau] \leq C \inf_{\tau\in\mathcal{T}_0} \mathbb{E}[(X^T\delta_0)^2|Q = \tau]$ for some constant $C < \infty$.

Remark 3.2. Assumptions 3.3 (iv) and 3.4 (ii) together imply that for all $\tau \in \mathcal{T}_0$,

$$\begin{aligned}\Delta_1(\tau) &\equiv \mathbb{E}\left[\left(\rho\left(Y, X^T\theta_0\right) - \rho\left(Y, X^T\beta_0\right)\right) 1\left\{\tau < Q \leq \tau_0\right\}\right] \geq c_0^2 \mathbb{P}\left[\tau < Q \leq \tau_0\right], \\ \Delta_2(\tau) &\equiv \mathbb{E}\left[\left(\rho\left(Y, X^T\beta_0\right) - \rho\left(Y, X^T\theta_0\right)\right) 1\left\{\tau_0 < Q \leq \tau\right\}\right] \geq c_0^2 \mathbb{P}\left[\tau_0 < Q \leq \tau\right].\end{aligned} \tag{3.2}$$

Note that Assumption 3.3 (ii) implies that $\Delta_1(\tau)$ is monotonely non-increasing when $\tau < \tau_0$, and $\Delta_2(\tau)$ is monotonely non-decreasing when $\tau > \tau_0$, respectively. Therefore, Assumptions 3.3 (ii), 3.3 (iv) and 3.4 (ii) all together imply that (3.2) holds for all τ in the $\mathcal{T}$, not just in the $\mathcal{T}_0$ since $\mathcal{T}$ is compact. Equation (3.2) plays an important role in achieving a super-efficient convergence rate for τ_0.

The following additional assumptions are useful to derive asymptotic results when $\delta_0 \neq 0$.

Assumption 3.5 (Moment bounds). *(i) There exist $0 < C_1 \leq C_2 < 1$ such that for all $\beta \in \mathbb{R}^p$ satisfying $\mathbb{E}|X^T\beta| \neq 0$,*

$$C_1 \leq \frac{\mathbb{E}[|X^T\beta| 1\{Q > \tau_0\}]}{\mathbb{E}|X^T\beta|} \leq C_2.$$

(ii) There exist constants $M > 0$ and $r > 0$ and the neighborhood $\mathcal{T}_0$ of τ_0 such that

$$\begin{aligned}\mathbb{E}\left[(X^T[(\theta - \beta) - (\theta_0 - \beta_0)])^2 \middle| Q = \tau\right] &\leq M, \\ \mathbb{E}[|X^T(\beta - \beta_0)| \big| Q = \tau] &\leq M, \\ \mathbb{E}[|X^T(\theta - \theta_0)| \big| Q = \tau] &\leq M, \\ \sup_{\tau \in \mathcal{T}_0 : \tau > \tau_0} \mathbb{E}\left[|X^T(\beta - \beta_0)| \frac{1\{\tau_0 < Q \leq \tau\}}{(\tau - \tau_0)}\right] &\leq M\mathbb{E}[|X^T(\beta - \beta_0)| 1\{Q \leq \tau_0\}], \\ \sup_{\tau \in \mathcal{T}_0 : \tau < \tau_0} \mathbb{E}\left[|X^T(\theta - \theta_0)| \frac{1\{\tau < Q \leq \tau_0\}}{(\tau_0 - \tau)}\right] &\leq M\mathbb{E}[|X^T(\theta - \theta_0)| 1\{Q > \tau_0\}],\end{aligned}$$

uniformly in $\beta \in \mathcal{B}(\beta_0, r)$, $\theta \in \mathcal{G}(\theta_0, r)$ and $\tau \in \mathcal{T}_0$.

Remark 3.3. Condition (i) requires that Q have non-negligible support on both sides of τ_0. Note that it is equivalent to

$$\begin{aligned}\left(\frac{1}{C_2} - 1\right) \mathbb{E}[|X^T\beta| 1\left\{Q > \tau_0\right\}] &\leq \mathbb{E}\left[|X^T\beta| 1\left\{Q \leq \tau_0\right\}\right] \\ &\leq \left(\frac{1}{C_1} - 1\right) \mathbb{E}|X^T\beta| 1\left\{Q > \tau_0\right\}.\end{aligned} \tag{3.3}$$

Hence this assumption prevents the conditional expectation of $X^T\beta$ given Q from changing too dramatically across regimes. Condition (ii) requires the boundedness and certain smoothness of the conditional expectation functions $\mathbb{E}[(X^T[(\theta-\beta)-(\theta_0-\beta_0)])^2\big|Q=\tau]$, $\mathbb{E}[|X^T(\beta-\beta_0)|\big|Q=\tau]$, and $\mathbb{E}[|X^T(\theta-\theta_0)|\big|Q=\tau]$, and prohibits degeneracy in one regime. The last two inequalities in condition (ii) are satisfied if

$$\frac{\mathbb{E}\left[\left|X^T\beta\right| |Q=\tau\right]}{\mathbb{E}\left[|X^T\beta|\right]} \le M$$

for all $\tau \in \mathcal{T}_0$ and for all β satisfying $0 < \mathbb{E}\left|X^T\beta\right| \le c$ for some small $c > 0$. In this view, we may regard (ii) as a local version of (i).

3.2 Risk consistency

Given the loss function $\rho(t_1, t_2)$, define the *excess risk* to be

$$R(\alpha,\tau) \equiv \mathbb{E}\rho(Y, X(\tau)^T\alpha) - \mathbb{E}\rho(Y, X(\tau_0)^T\alpha_0). \tag{3.4}$$

By the definition of (α_0, τ_0) in (2.2), we have that $R(\alpha,\tau) \ge 0$ for any $\alpha \in \mathcal{A}$ and $\tau \in \mathcal{T}$. The risk consistency is concerned about the convergence of $R(\widehat{\alpha},\widehat{\tau})$.

For a given sparse vector v with v_j indicating the j-th element of j, recall that its index set of nonzero components is defined by $J(v) \equiv \{j : v_j \ne 0\}$. Recall that the sparse coefficients in the two sub-populations $\{Q \le \tau_0\}$ and $\{Q > \tau_0\}$ are respectively β_0, θ_0, whose nonzero index sets are $J(\beta_0)$ and $J(\theta_0)$. Note that $J(\beta_0)$ and $J(\theta_0)$ can be different, admitting structural changes in the sparsity. Moreover, recall that the penalized M-estimation directly estimates $\alpha_0 = (\beta_0^T, \delta_0^T)^T$, with $\delta_0 = \theta_0 - \beta_0$. Hence the index set of all the nonzero components in the parameters is given by $J(\alpha_0)$.

In what follows, we denote $s = |J(\alpha_0)|_0$, as the cardinality of $J(\alpha_0)$. We allow that $s \to \infty$ as $n \to \infty$ and will give precise regularity conditions regarding its growth rates. The following result provides the consistency of the Lasso estimator in terms of the excess risk.

Theorem 3.1 (Risk consistency). *Let*

$$\omega_n \equiv (\log p)(\log n)\sqrt{\frac{\log p}{n}}. \tag{3.5}$$

Let Assumptions 3.1 (i)-(iii), 3.2 (i), 3.3 (i), and 3.4 (i) hold. Then, there exists some constant $C > 0$ such that for $\lambda_n = C\omega_n$,

$$R(\widehat{\alpha},\widehat{\tau}) = O_P(\omega_n s).$$

Theorem 3.1 shows the risk consistency if $\omega_n s \to 0$ as $n \to \infty$. The restriction on s is slightly stronger than that of the standard result $s = o(\sqrt{n/\log p})$ in the literature for the M-estimation (see, e.g. Bühlmann and van de Geer (2011), Chapter 6.6). Our situation is also different from the setup studied by van de Geer (2008) since the objective function $\rho(Y, X(\tau)^T\alpha)$ is non-convex in τ, due to the unknown change-point. The extra logarithmic factor $(\log p)(\log n)$ is due to the existence of the unknown and possibly non-identifiable threshold parameter τ_0. In fact, an inspection of the proof of Theorem 3.1 reveals that it suffices to assume that ω_n satisfies $\omega_n \gg \log_2(p/s)[\log(np)/n]^{1/2}$. The term $\log_2(p/s)$ and the additional $(\log n)^{1/2}$ term inside the brackets are needed to establish the stochastic continuity of the empirical process

$$\nu_n(\alpha, \tau) \equiv \frac{1}{n}\sum_{i=1}^{n}\left[\rho\left(Y_i, X_i(\tau)^T\alpha\right) - \mathbb{E}\rho\left(Y, X(\tau)^T\alpha\right)\right].$$

uniformly over $(\alpha, \tau) \in \mathcal{A} \times \mathcal{T}$.

3.3 Threshold consistency

In this subsection, we establish conditions under which the unknown change-point τ_0 is identifiable, and thus it can be consistently estimated. In addition, we present theoretical analysis when τ_0 is non-identifiable in Section 3.5. Intuitively, if there is no structural change in the sparsity, $\delta_0 = 0$, then $\rho(Y, X^T\beta_0 + X^T\delta_0 1\{Q > \tau_0\})$ will be observationally equivalent regardless of the value of $\tau_0 \in \mathcal{T}$, and in that case it is impossible to consistently estimate τ_0. As a result, τ_0 is identifiable only if δ_0 is "significantly" nonzero, which leads to a structural change.

Theorem 3.2 (Consistency of $\widehat{\tau}$)**.** *Let Assumptions 3.1 (i)-(iii), 3.2 (i)-(ii), 3.3 (i)-(iv), 3.4 (i)-(ii), and 3.5 (i) hold. Then,* $\widehat{\tau} \xrightarrow{p} \tau_0$.

We briefly provide the logic behind the proof of Theorem 3.2 here. Note that for all $\alpha \equiv (\beta^T, \delta^T)^T \in \mathbb{R}^{2p}$ and $\theta \equiv \beta + \delta$, the excess risk has the following decomposition: when $\tau_1 < \tau_0$,

$$\begin{aligned} R(\alpha, \tau_1) = {} & \mathbb{E}\left(\left[\rho\left(Y, X^T\beta\right) - \rho\left(Y, X^T\beta_0\right)\right] 1\{Q \le \tau_1\}\right) \\ & + \mathbb{E}\left(\left[\rho\left(Y, X^T\theta\right) - \rho\left(Y, X^T\theta_0\right)\right] 1\{Q > \tau_0\}\right) \\ & + \mathbb{E}\left(\left[\rho\left(Y, X^T\theta\right) - \rho\left(Y, X^T\beta_0\right)\right] 1\{\tau_1 < Q \le \tau_0\}\right), \end{aligned} \tag{3.6}$$

and when $\tau_2 > \tau_0$,

$$\begin{aligned} R(\alpha, \tau_2) = & \mathbb{E}\left(\left[\rho\left(Y, X^T\beta\right) - \rho\left(Y, X^T\beta_0\right)\right] 1\{Q \le \tau_0\}\right) \\ & + \mathbb{E}\left(\left[\rho\left(Y, X^T\theta\right) - \rho\left(Y, X^T\theta_0\right)\right] 1\{Q > \tau_2\}\right) \\ & + \mathbb{E}\left(\left[\rho\left(Y, X^T\beta\right) - \rho\left(Y, X^T\theta_0\right)\right] 1\{\tau_0 < Q \le \tau_2\}\right). \end{aligned} \tag{3.7}$$

The key observations are that all the six terms in the above decompositions are non-negative, and are stochastically negligible when taking $\alpha = \widehat{\alpha}$, and $\tau_1 = \widehat{\tau}$ if $\widehat{\tau} < \tau_0$ or $\tau_2 = \widehat{\tau}$ if $\widehat{\tau} > \tau_0$. This follows from the risk consistency of $R(\widehat{\alpha}, \widehat{\tau})$. Then, the identification conditions for α_0 and τ_0 (Assumptions 3.3 (ii)-(iv)), along with Assumption 3.5 (i), are useful to show that the risk consistency implies the consistency of $\widehat{\tau}$.

3.4 Rate of convergence and super-efficiency when τ_0 is identifiable

This subsection derives the rate of convergence for the excess risk as well as $|\widehat{\alpha} - \alpha_0|_1$, and proves that we can achieve the super-convergence rate for $\widehat{\tau} - \tau_0$ when τ_0 is identifiable.

We first make an assumption that is an extension of the well-known *compatibility condition* (see Bühlmann and van de Geer (2011), Chapter 6), which is related to the "restricted eigenvalue condition" of Bickel et al. (2009). Both conditions are commonly assumed in high-dimensional sparse literature. See, e.g. van de Geer and Bühlmann (2009) for the relations among these conditions on the design matrix. In particular, the following condition is a uniform-in-τ version of the compatibility condition.

For a $2p$ dimensional vector α, we shall use α_J and α_{J^c} to denote its subvectors formed by indices in $J(\alpha_0)$ and $\{1, ..., 2p\}/J(\alpha_0)$, respectively.

Assumption 3.6 (Compatibility condition)**.** *There is a neighborhood $\mathcal{T}_0 \subset \mathcal{T}$ of τ_0, and a constant $\phi > 0$ such that for all $\tau \in \mathcal{T}_0$ and all $\alpha \in \mathbb{R}^{2p}$ satisfying $|\alpha_{J^c}|_1 \le 5|\alpha_J|_1$,*

$$\phi|\alpha_J|_1^2 \le s\alpha^T \mathbb{E}[X(\tau)X(\tau)^T]\alpha. \tag{3.8}$$

Note that Assumption 3.6 requires that the compatibility condition hold uniformly in τ over a small neighbourhood of τ_0. Note that Assumption 3.6 is imposed on the population covariance matrix $\mathbb{E}[X(\tau)X(\tau)^T]$, so a simple sufficient condition of Assumption 3.6 is that the smallest eigenvalue of $\mathbb{E}[X(\tau)X(\tau)^T]$ is bounded away from zero uniformly in $\tau \in \mathcal{T}_0$. Even if $p > n$, the population covariance can still be strictly positive definite while the sample covariance $\frac{1}{n}\sum_{i=1}^n X_i(\tau)X_i(\tau)^T$ is not.

Example 3.1 (Factor analysis with structural-changing loadings)**.** Suppose the regressors satisfy a factor-structure with a change point:

$$X_i(\tau) = \Lambda(\tau) f_i + u_i, \quad i = 1, ..., n,$$

where $\Lambda(\tau)$ is a $2p \times k$ dimensional loading matrix that may depend on the change point τ, and f_i is a k-dimensional vector of common factors that may not be observable; u_i is the error term for factor analysis that is independent of f_i. Let $\mathrm{cov}(f_i)$ denote the $k \times k$ covariance of f_i. Then the covariance of the random design matrix has the following decomposition:

$$\mathbb{E}[X_i(\tau)X_i(\tau)^T] = \Lambda(\tau)\mathrm{cov}(f_i)\Lambda(\tau)^T + \mathbb{E}[u_i u_i^T].$$

Then a sufficient condition of Assumption 3.6 is that all the eigenvalue of $\mathbb{E}[u_i u_i^T]$ are bounded below by a constant $c_{\min}$, and (3.8) is satisfied for $\phi(J) = c_{\min}$. Note that assuming the minimum eigenvalue of $\mathbb{E}[u_i u_i^T]$ to be bounded below is not stringent, because for the identifiability purpose, $\mathbb{E}[u_i u_i^T]$ is often assumed to be diagonal (e.g., Lawley and Maxwell (1971)). Then it is sufficient to have $\min_{j \le 2p} \mathbb{E}[u_{ij}^2] > c_{\min}$. □

Remark 3.4. In high-dimensional M-estimation, it is necessary to impose a version of margin condition in addition to the compatibility condition, when viewing the sparse recovery as a type of inverse problem. Note that the margin condition, together with the compatibility condition, are sufficient to the so-called *restricted strong convexity* condition in Negahban et al. (2012) and Loh and Wainwright (2013). Recall that Assumptions 3.3 (iii) and (iv) are the corresponding conditions for α_0 and τ_0, respectively.

The following theorem presents the rates of convergence. Recall

$$\omega_n = (\log p)(\log n)\sqrt{\frac{\log p}{n}}, \quad \text{and } s = |J(\alpha_0)|_0. \tag{3.9}$$

Theorem 3.3 (Rates of convergence)**.** *Suppose that $\omega_n s^2 \log p = o(1)$. Let $\lambda_n = C\omega_n$ for some constant $C > 0$. Then under Assumptions 3.1-3.6, we have:*

$$|\widehat{\alpha} - \alpha_0|_1 = O_P(\omega_n s), \; R(\widehat{\alpha}, \widehat{\tau}) = O_P(\omega_n^2 s), \quad \textit{and} \quad |\widehat{\tau} - \tau_0| = O_P(\omega_n^2 s/\Delta_0),$$

where $\Delta_0 \equiv \inf_{\tau \in \mathcal{T}_0} \mathbb{E}[(X^T\delta_0)^2 | Q = \tau]$.

The required growth rate on the sparsity index s can be rewritten as $s^4(\log p)^5(\log n)^2 = o(n)$. The achieved convergence rate for $\widehat{\alpha}$ is slightly slower than the usual rate for Lasso estimation (e.g., Bickel et al. (2009)), with an additional factor $(\log p)(\log n)$, due to the unknown change-point τ_0. In Section 4, we will see that the rate of convergence for $\widehat{\alpha}$ can be improved (via a second-step regularization) to be the oracle rate of convergence when signals are relatively strong.

Remark 3.5. It is worth noting that the convergence rate of $\widehat{\tau}$ depends on Δ_0, which is assumed to be bounded or to diverge to infinity at the rate of $\Delta_0 = O(|\delta_0|_2^2)$ (see Assumption 3.4 (iii)). Moreover, note that $\widehat{\tau}$ converges to τ_0 at least as fast as $R(\widehat{\alpha}, \widehat{\tau})$ and its rate of convergence can be faster than the standard parametric rate of $n^{-1/2}$, as long as $s^2(\log p)^6(\log n)^4/\Delta_0 = o(n)$. The main reason we achieve the super-consistency for estimating τ_0 is that our objective function behaves locally linearly around τ_0 with a kink at τ_0 (granted in Remark 3.2), unlike in the regular estimation problem where an objective function behaves locally quadratically around the true parameter value.

3.5 Rate of convergence when τ_0 is non-identifiable

In this subsection, we derive the rate of convergence for the excess risk as well as $|\widehat{\alpha} - \alpha_0|_1$ when there is no structural change in the sparsity (that is, $\delta_0 = 0$). Recall that when $\delta_0 = 0$,

$$X^T\beta_0 + X^T\delta_0 1\{Q > \tau_0\} = X^T\beta_0,$$

hence τ_0 is non-identifiable.

Most of the required conditions are similar as before, except that the identification and margin conditions with respect to τ_0 are not required. We shall see below that the required conditions are slightly stronger than those of the regular ℓ_1-penalized regressions (e.g., Bühlmann and van de Geer (2011), Chapter 6.6), as we do not know whether any structural changes on the sparsity are present. Some conditions are required to be valid uniformly over the parameter space for τ.

Specifically, define

$$\begin{aligned}
\tilde{\mathcal{B}}(\beta_0, r, \tau) &= \{\beta \in \mathcal{B} \subset \mathbb{R}^p : \mathbb{E}[(X^T(\beta - \beta_0))^2 1\{Q \leq \tau\}] \leq r^2\}, \\
\tilde{\mathcal{G}}(\beta_0, r, \tau) &= \{\theta \in \mathcal{G} \subset \mathbb{R}^p : \mathbb{E}[(X^T(\theta - \beta_0))^2 1\{Q > \tau\}] \leq r^2\}.
\end{aligned} \tag{3.10}$$

For a constant $\eta > 0$, define

$$\begin{aligned}
\tilde{r}_1(\eta) \equiv \sup_r \Big\{ r : &\ \mathbb{E}\left(\left[\rho\left(Y, X^T\beta\right) - \rho\left(Y, X^T\beta_0\right)\right] 1\{Q \leq \tau\}\right) \\
&\geq \eta \mathbb{E}[(X^T(\beta - \beta_0))^2 1\{Q \leq \tau\}] \text{ for all } \beta \in \tilde{\mathcal{B}}(\beta_0, r, \tau) \text{ and for all } \tau \in \mathcal{T} \Big\}
\end{aligned}$$

and

$$\tilde{r}_2(\eta) \equiv \sup_r \Big\{ r : \mathbb{E}\left(\left[\rho\left(Y, X^T\theta\right) - \rho\left(Y, X^T\beta_0\right)\right] 1\left\{Q > \tau\right\}\right) \\ \geq \eta \mathbb{E}[(X^T(\theta - \beta_0))^2 1\{Q > \tau\}] \text{ for all } \theta \in \tilde{\mathcal{G}}(\beta_0, r, \tau) \text{ and for all } \tau \in \mathcal{T} \Big\}.$$

Note that in this case β_0 is the vector of true regression coefficients, so we have the following assumption for identification.

Assumption 3.7 (Identification of α_0 when $\delta_0 = 0$). *(i) For all $\alpha \in \mathcal{A}$ and for all $\tau \in \mathcal{T}$, almost surely,*

$$\mathbb{E}[\rho(Y, X(\tau)^T\alpha) - \rho(Y, X^T\beta_0)|Q] \geq 0,$$

(ii) There exist constants $\eta^* > 0$ and $r^* > 0$ such that $\tilde{r}_1(\eta^*) \geq r^*$ and $\tilde{r}_2(\eta^*) \geq r^*$.

Assumption 3.7 is a uniform-in-τ ($\in \mathcal{T}$) version of Assumptions 3.3 (ii) and (iii) since τ_0 is non-identifiable when $\delta_0 = 0$. Similarly, we impose below the compatibility condition uniformly over $\mathcal{T}$. In this case $J(\alpha_0) = J(\beta_0)$, so $\alpha_J = (\alpha_j : j \in J(\beta_0))$.

Assumption 3.8 (Compatibility condition when $\delta_0 = 0$). *There is a constant $\phi > 0$ such that for all $\tau \in \mathcal{T}$ and all $\alpha \in \mathbb{R}^{2p}$ satisfying $|\alpha_{J^c}|_1 \leq 4|\alpha_J|_1$,*

$$\phi|\alpha_J|_1^2 \leq s\alpha^T \mathbb{E}[X(\tau)X(\tau)^T]\alpha. \tag{3.11}$$

The following theorem gives the rates of convergence for $|\widehat{\alpha} - \alpha_0|_1$ and $R(\widehat{\alpha}, \widehat{\tau})$ when $\delta_0 = 0$. Recall

$$\omega_n = (\log p)(\log n)\sqrt{\frac{\log p}{n}}, \quad \text{and } s = |J(\beta_0)|_0.$$

Theorem 3.4 (Rates of convergence when $\delta_0 = 0$). *Suppose $\omega_n s = o(1)$. Let $\lambda_n = C\omega_n$ for some constant $C > 0$. Then under Assumptions 3.1 (i)-(iii), 3.2 (i), 3.3 (i), 3.4 (i), 3.7, and 3.8, we have that*

$$|\widehat{\alpha} - \alpha_0|_1 = O_P(\omega_n s) \quad \textit{and} \quad R(\widehat{\alpha}, \widehat{\tau}) = O_P(\omega_n^2 s).$$

The results obtained in Theorem 3.4 combined with those obtained in Theorem 3.3 imply that the Lasso estimator performs equally well in terms of both the ℓ_1 loss for $\widehat{\alpha}$ and the excess risk, regardless of the existence of the threshold effect. This type of the advantage of the ℓ_1-penalized estimator was established in Lee et al. (2014) for the Gaussian mean regression model with a deterministic design.

Remark 3.6. Note that when the model does not have any structural change, our rate of convergence is slightly slower than the usual ℓ_1-penalized regression for M-estimation (e.g., van de Geer (2008) and Negahban et al. (2012)). The additional factor $(\log p)(\log n)$ is due to the fact that $\delta_0 = 0$ is unknown, and arises from technical arguments that bound empirical processes uniformly over τ.

4 Theoretical Properties of the SCAD Estimator

4.1 Oracle Properties of $\widetilde{\alpha}$

While we allow the loss function $\rho(y,t)$ to be non-differentiable such that models like quantile regressions can be included, it is often the case that the differentiability holds after expectations are taken. We require that $\mathbb{E}[\rho(Y, X(\tau)^T\alpha)]$ be differentiable with respect to α, and define

$$m_j(\tau,\alpha) \equiv \frac{\partial \mathbb{E}[\rho(Y, X(\tau)^T\alpha)]}{\partial \alpha_j}, \quad m(\tau,\alpha) \equiv (m_1(\tau,\alpha), ..., m_{2p}(\tau,\alpha))^T.$$

Also, let $m_J(\tau,\alpha) \equiv (m_j(\tau,\alpha) : j \in J(\alpha_0))$. Assume in the section that α_0 is in the interior of the parameter space $\mathcal{A}$. Hence, we have that $m(\tau_0,\alpha_0) = 0$.

When the assumptions of Section 3.4 are satisfied so that τ_0 is identifiable, we impose the following conditions.

Assumption 4.1 (Conditions for the population objective function)**.** *$\mathbb{E}[\rho(Y, X(\tau)^T\alpha)]$ is three times continuously differentiable with respect to α, and there are constants $c_1, c_2, L > 0$ and a neighborhood $\mathcal{T}_0$ of τ_0 such that the following conditions hold: for all large n and all $\tau \in \mathcal{T}_0$,*

(i) There is $M_n > 0$, which may depend on the sample size, such that

$$\max_{j\le 2p} |m_j(\tau,\alpha_0) - m_j(\tau_0,\alpha_0)| < M_n|\tau - \tau_0|.$$

(ii) There is $r > 0$ such that for all $\beta \in \mathcal{B}(\beta_0, r)$, $\theta \in \mathcal{G}(\theta_0, r)$, $\alpha = (\beta^T, \theta^T - \beta^T)^T$ satisfies:

$$\max_{j\le 2p} \sup_{\tau\in\mathcal{T}_0} |m_j(\tau,\alpha) - m_j(\tau,\alpha_0)| < L\,|\alpha - \alpha_0|_1\,.$$

(iii) α_0 *is in the interior of the parameter space* $\mathcal{A}$*, and*

$$\inf_{\tau \in \mathcal{T}_0} \lambda_{\min} \left(\frac{\partial^2 \mathbb{E}[\rho(Y, X_J(\tau)^T \alpha_{0J})]}{\partial \alpha_J \partial \alpha_J^T} \right) > c_1.$$

$$\sup_{|\alpha_J - \alpha_{0J}|_1 < c_2,} \sup_{\tau \in \mathcal{T}_0} \max_{i,j,k \in J} \left| \frac{\partial^3 \mathbb{E}[\rho(Y, X_J(\tau)^T \alpha_J)]}{\partial \alpha_i \partial \alpha_j \partial \alpha_k} \right| < L.$$

The score-condition in the population level is expressed by $m(\tau_0, \alpha_0) = 0$ since α_0 is in the interior of $\mathcal{A}$ by condition (iii). Conditions (i) and (ii) regulate the continuity of the score $m(\tau, \alpha)$, and Condition (iii) assumes the higher-order differentiability of the loss function $\rho(y, t)$. Condition (i) requires the Lipschitz continuity of the score function with respect to the threshold. The Lipschitz constant may grow with n, since it is assumed uniformly over $j \leq 2p$. In many interesting examples being considered, M_n in fact grows slowly; as a result, it does not affect the asymptotic behavior of $\widetilde{\alpha}$. For the logistic and quantile regression models, we will show that $M_n = Cs^{1/2}$ for some constant $C > 0$. Condition (ii) requires the local equi-continuity at α_0 in the ℓ_1 norm of the class

$$\{m_j(\tau, \alpha) : \tau \in \mathcal{T}_0, j \leq 2p\}.$$

In Section 4, we shall verify Assumption 4.1 in both the logistic and quantile regression models. Under the foregoing assumptions, the following two theorems establish the oracle properties of the adaptively weighted-ℓ_1-regularized estimators.

We partition $\widetilde{\alpha} = (\widetilde{\alpha}_J, \widetilde{\alpha}_{J^c})$ so that $\widetilde{\alpha}_J = (\widetilde{\alpha}_j : j \in J(\alpha_0))$ and $\widetilde{\alpha}_{J^c} = (\widetilde{\alpha}_j : j \notin J(\alpha_0))$. Note that $\widetilde{\alpha}_J$ consists of the estimators of β_{0J} and δ_{0J}, whereas $\widetilde{\alpha}_{J^c}$ consists of the estimators of all the zero components of β_0 and δ_0. Let $\alpha_{0J}^{(j)}$ denote the j-th element of α_{0J}, where $j \in J(\alpha_0)$.

Theorem 4.1 (Oracle properties)**.** *Suppose that* $s^4(\log p)^3(\log n)^3 + sM_n^4(\log p)^6(\log n)^6 = o(n)$*, Assumptions 3.1-3.6, 4.1 are satisfied, and*

$$\omega_n + s\sqrt{\frac{\log s}{n}} + M_n \omega_n^2 s \log n \ll \mu_n \ll \min_{j \in J(\alpha_0)} |\alpha_{0J}^{(j)}|.$$

Then

$$|\widetilde{\alpha}_J - \alpha_{0J}|_2 = O_P\left(\sqrt{\frac{s \log s}{n}}\right), \quad |\widetilde{\alpha}_J - \alpha_{0J}|_1 = O_P\left(s\sqrt{\frac{\log s}{n}}\right)$$

and

$$P(\widetilde{\alpha}_{J^c} = 0) \to 1.$$

The required condition $s^4(\log p)^3(\log n)^3 + sM_n^4(\log p)^6(\log n)^6 = o(n)$ is the price paid for not knowing τ_0. Under this condition, the effect of estimating τ_0 is negligible. Note that $\omega_n \ll \mu_n$. It is known that the variable selection consistency often requires a larger tuning parameter than does prediction (e.g., Sun and Zhang (2012)). Some additional remarks are in order.

Remark 4.1. Variable selection consistency often requires the minimal signal to be clearly separated from zero. We note that the required signal strength $\min_{j\in J(\alpha_0)} |\alpha_{0J}^{(j)}|$ is stronger than that of the existing literature without a structural change in the sparsity. This is natural, since with an unknown change point, the noise level is higher, and additional estimation errors coming from estimating τ_0 needs to be taken into account. On the other hand, it is assuring to see that our estimation consistency results achieved in Section 3 do not require this kind of conditions.

Remark 4.2. We have achieved the fast rates of convergence $O_P(s\sqrt{\log s/n})$ and $O_P(\sqrt{s\log s/n})$ in the ℓ_1 and ℓ_2 distances, respectively. Compared to the convergence rate in Theorem 3.3, we see that after the consistent variable selection, the ℓ_1 rate of convergence is improved, and the ℓ_2 rate is slightly faster than the sparse minimax ℓ_2 rate $O_P(\sqrt{s\log p/n})$ in, e.g., Johnstone (1994) and Raskutti et al. (2010), which is a natural result. Intuitively, as $\widehat{J}$ consistently recovers $J(\alpha_0)$, the price $\sqrt{\log p}$ for not knowing $J(\alpha_0)$ can be avoided.

We now consider the case when $\delta_0 = 0$. In this case, τ_0 is not identifiable, and there is actually no structural change in the sparsity. If α_0 is in the interior of $\mathcal{A}$, then $m(\tau, \alpha_0) = 0$ for all $\tau \in \mathcal{T}$, and Assumption 4.1 is revised as follows.

Assumption 4.2 (Conditions when $\delta_0 = 0$). *$\mathbb{E}[\rho(Y, X(\tau)^T\alpha)]$ is three times differentiable with respect to α, and there are constants $c_1, c_2, L > 0$ such that when $\delta_0 = 0$, for all large n,*
(i) There is $r > 0$ such that for all $\beta \in \mathcal{B}(\beta_0, r)$, $\theta \in \mathcal{G}(\theta_0, r)$, $\alpha = (\beta^T, \theta^T - \beta^T)^T$ satisfies:

$$\max_{j\le 2p} \sup_{\tau\in\mathcal{T}} |m_j(\tau,\alpha) - m_j(\tau,\alpha_0)| < L\,|\alpha-\alpha_0|_1 .$$

(ii) α_0 is in the interior of the parameter space $\mathcal{A}$, and

$$\lambda_{\min}\left(\frac{\partial^2 \mathbb{E}[\rho(Y, X_{J(\beta_0)}^T\beta_{0J})]}{\partial\beta_J\partial\beta_J^T}\right) > c_1.$$

$$\sup_{|\alpha_J-\alpha_{0J}|_1<c_2,}\ \max_{i,j,k\in J(\beta_0)} \left|\frac{\partial^3\mathbb{E}[\rho(Y, X_{J(\beta_0)}^T\beta_J)]}{\partial\beta_i\partial\beta_j\partial\beta_k}\right| < L.$$

Below we write $\widetilde{\alpha} = (\widetilde{\beta}^T, \widetilde{\delta}^T)^T$, and write $\widetilde{\beta}_J = (\widetilde{\beta}_j : j \in J(\beta_0))$, $\widetilde{\beta}_{J^c} = (\widetilde{\beta}_j : j \notin J(\beta_0))$.

Theorem 4.2 (Oracle properties when $\delta_0 = 0$)**.** *Consider the case where the true $\delta_0 = 0$. Suppose $s^4(\log s) = o(n)$, Assumptions 3.1 (i)-(iii), 3.2 (i), 3.3 (i), 3.4 (i), 3.7, 3.8 and 4.2 hold, and*

$$\omega_n + s\sqrt{\frac{\log s}{n}} + M_n \omega_n^2 s \log n \ll \mu_n \ll \min_{j \in J(\alpha_0)} |\alpha_{0J}^{(j)}|.$$

Then

$$\left|\widetilde{\beta}_J - \beta_{0J}\right|_2 = O_P\left(\sqrt{\frac{s \log s}{n}}\right), \quad \left|\widetilde{\beta}_J - \beta_{0J}\right|_1 = O_P\left(s\sqrt{\frac{\log s}{n}}\right),$$

and

$$P(\widetilde{\beta}_{J^c} = 0) \to 1, \quad P(\widetilde{\delta} = 0) \to 1.$$

Interestingly, Theorem 4.2 demonstrates that when there is in fact no structural change in the sparsity, our estimator for δ_0 is exactly zero with a high probability. Therefore, the estimator can also be used as a diagnostic tool to check whether any structural changes are present.

4.2 Asymptotic Distribution for $\widetilde{\alpha}_J$ and $\widetilde{\tau}$

Thanks to the variable selection consistency established in the previous section, it suffices to consider

$$\operatorname*{argmin}_{\alpha_J} \frac{1}{n} \sum_{i=1}^{n} \rho(Y_i, X_{iJ}(\widehat{\tau})^T \alpha_J),$$

where α_J is a subvector of α projected on the oracle space $J(\alpha_0)$. Recall that by Theorems 3.3 and 4.1, we have that

$$|\widetilde{\alpha}_J - \alpha_{0J}|_2 = O_P\left(\sqrt{\frac{s \log s}{n}}\right) \quad \text{and} \quad |\widehat{\tau} - \tau_0| = O_P\left[(\log p)^3 (\log n)^2 \frac{s}{n}\right]. \tag{4.1}$$

In view of (4.1), define $r_n \equiv \sqrt{n^{-1} s \log s}$ and $s_n \equiv n^{-1} s[(\log p)^3 (\log n)^2]$. Let

$$Q_n^*(\alpha_J, \tau) \equiv \frac{1}{n} \sum_{i=1}^{n} \rho\left(Y_i, X_{iJ}(\tau)^T \alpha_J\right),$$

where $\alpha_J \in \mathcal{A}_n \equiv \{\alpha_J : |\alpha_J - \alpha_{0J}|_2 \le K r_n\} \subset \mathbb{R}^s$ and $\tau \in \mathcal{T}_n \equiv \{\tau : |\tau - \tau_0| \le K s_n\}$ for some $K < \infty$, where K is a generic finite constant.

The following lemma is useful to establish that α_0 can be estimated as if τ_0 were known and vice versa.

Lemma 4.1 (Asymptotic Equivalence)**.** *Assume that* $\frac{\partial}{\partial\alpha}E\left[\rho\left(Y, X^T\alpha\right)|Q=t\right]$ *exists for all* t *in a neighborhood of* τ_0 *and all its elements are continuous and bounded below and above. Suppose that* $s^5(\log s)(\log p)^6(\log n)^4 = o\left(n\right)$*. Then*

$$\sup_{\alpha_J\in\mathcal{A}_n,\tau\in\mathcal{T}_n}\left|\{Q_n^*\left(\alpha_J,\tau\right)-Q_n^*\left(\alpha_J,\tau_0\right)\}-\{Q_n^*\left(\alpha_{0J},\tau\right)-Q_n^*\left(\alpha_{0J},\tau_0\right)\}\right| = o_P\left(n^{-1}\right).$$

This lemma implies that the asymptotic distribution of $\widehat{\alpha}_J \equiv \operatorname{argmin}_{\alpha_J} Q_n^*\left(\alpha_J,\widehat{\tau}\right)$ can be characterized by $\widehat{\alpha}_J^* \equiv \operatorname{argmin}_{\alpha_J} Q_n^*\left(\alpha_J,\tau_0\right)$ for any estimator $\widehat{\tau}$ with a convergence rate at least s_n. Furthermore, the variable selection consistency implies that the asymptotic distribution of the SCAD estimator $\widetilde{\alpha}_J$ is equivalent to that of $\widehat{\alpha}_J^*$. Then following the existing results on M-estimation with parameters of increasing dimension (see, e.g. He and Shao (2000)), the asymptotic normality of a linear transformation of $\widetilde{\alpha}_J$, i.e., $R\widetilde{\alpha}_J$, where $R:\mathbb{R}^s\to\mathbb{R}$ with $|R|_2=1$, can be established.

Remark 4.3. The asymptotic normality of $\widetilde{\alpha}_J$ reveals the oracle asymptotic behavior of the estimator in two senses: (1) it is the same distribution as that of the estimate restricted on the oracle set J, thereby implying that the asymptotic normality can be established regardless of $\delta_0 = 0$, and (2) the effect of estimating τ_0 is asymptotically negligible when τ_0 is identifiable. Hence the limiting distribution is also the same as if τ_0 were known *a priori*. The first phenomenon is mainly due to the variable selection consistency and the use of the asymptotic unbiased penalty (SCAD-weighted-ℓ_1), while the second phenomenon is mainly due to the super-efficiency for estimating τ_0 (the fast rate of convergence, as in Theorem 3.3). Consequently, there is no first-order efficiency loss.

We now discuss asymptotic properties of $\widetilde{\tau}$. Recall that

$$\widetilde{\tau} = \operatorname*{argmin}_{\tau}\frac{1}{n}\sum_{i=1}^{n}\rho\left(Y_i, X_i\left(\tau\right)^T\widetilde{\alpha}\right),$$

where $\widetilde{\alpha}$ is the second step SCAD estimate of α. Due to the selection consistency of the SCAD, i.e. $\widetilde{\alpha}_{J^C} = 0$ with probability approaching one, $\widetilde{\tau}$ is equivalent, with the same probability, to

$$\operatorname*{argmin}_{\tau}\frac{1}{n}\sum_{i=1}^{n}\rho\left(Y_i, X_{Ji}\left(\tau\right)^T\widetilde{\alpha}_J\right).$$

Suppose that $\delta_0 \neq 0$. For brevity, we focus on the case that Δ_0 is bounded. Then, in view of Lemma 4.1 and the n-consistency of the threshold estimate in the standard threshold models with the fixed small number of regressors, we can conclude that $n\left(\tilde{\tau}-\tau_0\right)$ is asymptotically equivalent to

$$\underset{|h|\leq K}{\operatorname{argmax}} - \left(\sum_{i=1}^{n} \rho_i 1\left\{\tau_0 < Q_i \leq \tau_0 + hn^{-1}\right\} - \sum_{i=1}^{n} \rho_i 1\left\{\tau_0 + hn^{-1} < Q_i \leq \tau_0\right\} \right) \text{ for some } K < \infty,$$

where $\rho_i = \rho\left(Y_i, X_i^T\beta_0\right) - \rho\left(Y_i, X_i^T\beta_0 + X_i^T\delta_0\right)$. The weak convergence of this process for a variety of M-estimators is well known in the literature (see e.g. Pons (2003); Kosorok and Song (2007); Lee and Seo (2008)) and the argmax continuous mapping theorem by Seijo and Sen (2011b) yields the asymptotic distribution, namely the smallest maximizer of a compound Poisson process.

5 Applications to Quantile and Logistic Regression Models

This section considers two important examples: quantile and logistic regression models. We present mild primitive conditions under which all the regularity conditions assumed in Sections 3 and 4 are satisfied. For brevity, we focus on the case $\delta_0 \neq 0$ here, since the primitive conditions under $\delta_0 = 0$ can be obtained similarly.

5.1 Quantile regression with a change point

The quantile regression with a change point is modeled as follows:

$$\begin{aligned} Y &= X^T\beta_0 + X^T\delta_0 1\{Q > \tau_0\} + U \\ &\equiv X(\tau_0)^T\alpha_0 + U, \end{aligned}$$

where the regression error U satisfies the conditional restriction $\mathbb{P}(U \leq 0|X, Q) = \gamma$ for some known $\gamma \in (0, 1)$. The rate of convergence of ℓ_1-penalized estimation for sparse quantile regression has been studied by Belloni and Chernozhukov (2011). Unlike the mean regressions, sparse quantile regression analyzes the effects of active regressors on different parts of the conditional distribution of a response variable. Since loss function of the quantile regression is non-smooth, it has been treated separately from the usual M-estimation framework. The literature also includes Wang (2013); Bradic et al. (2011); Wang et al. (2012) and Fan et al. (2014a) among others. All the aforementioned papers are under the homogeneous sparsity framework (equivalently, knowing that $\delta_0 = 0$ in the quantile regression model). Ciuperca (2013) considers penalized estimation of a quantile regression model with breaks, but the corresponding analysis is restricted to the case when the number of potential covariates is small, and is not about structural changes in sparsity.

The loss function for quantile regression is defined as

$$\rho(Y, X(\tau)^T\alpha) = (Y - X(\tau)^T\alpha)(\gamma - 1\{Y - X(\tau)^T\alpha \leq 0\}),$$

which is not differentiable in α. But it is straightforward to check that in this case

$$m_j(\tau, \alpha) = \mathbb{E}[X_j(\tau)(1\{Y - X(\tau)^T\alpha \leq 0\} - \gamma)],$$

and if the conditional distribution of $Y|(X,Q)$ has a bounded density function $f_{Y|X,Q}(y|X,Q)$, then

$$\frac{\partial^2 \mathbb{E}[\rho(Y, X_J(\tau)^T\alpha_{0J})]}{\partial\alpha_J\partial\alpha_J^T} = \mathbb{E}[X_J(\tau)X_J(\tau)^T f_{Y|X,Q}(X(\tau)^T\alpha_0|X,Q)] \equiv \Gamma(\tau, \alpha_0).$$

We make the following assumptions for quantile regression models.

Assumption 5.1. *(i) The conditional distribution $Y|X,Q$ has a continuously differentiable density function $f_{Y|X,Q}(y|x,q)$, whose derivative with respect to y is denoted by $\tilde{f}_{Y|X,Q}(y|x,q)$.*

(ii) There are constants $C_1, C_2 > 0$ such that that for all (y,x,q) in the support of (Y,X,Q),

$$|\tilde{f}_{Y|X,Q}(y|x,q)| \leq C_1, \quad f_{Y|X,Q}(x(\tau_0)^T\alpha_0|x,q) \geq C_2.$$

(iii) There exists a constant $r^*_{QR} > 0$ such that

$$\inf_{\beta\in\mathcal{B}(\beta_0, r^*_{QR}), \beta\neq\beta_0} \frac{\mathbb{E}[|X^T(\beta-\beta_0)|^2 1\{Q \leq \tau_0\}]^{3/2}}{\mathbb{E}[|X^T(\beta-\beta_0)|^3 1\{Q \leq \tau_0\}]} \geq r^*_{QR}\frac{2C_1}{3C_2} > 0 \tag{5.1}$$

and that

$$\inf_{\theta\in\mathcal{G}(\theta_0, r^*_{QR}), \theta\neq\theta_0} \frac{\mathbb{E}[|X^T(\theta-\theta_0)|^2 1\{Q > \tau_0\}]^{3/2}}{\mathbb{E}[|X^T(\theta-\theta_0)|^3 1\{Q > \tau_0\}]} \geq r^*_{QR}\frac{2C_1}{3C_2} > 0. \tag{5.2}$$

(iv) $\Gamma(\tau, \alpha_0)$ is positive definite uniformly in a neighborhood of τ_0.

Conditions (i) and (ii) are standard assumptions for quantile regression models. Condition (iii) is a kind of the "restricted nonlinearity" condition, similar to condition D.4 in Belloni and Chernozhukov (2011). Condition (iv) is a weak condition that imposes non-singularity of the Hessian matrix of the population objective function uniformly in a neighborhood of τ_0.

Remark 5.1. As pointed out by Belloni and Chernozhukov (2011, online supplement), If $X^T c$ follows a logconcave distribution conditional on Q for any nonzero c (e.g. if the distribution of X is multivariate normal), then Theorem 5.22 of Lovász and Vempala (2007) and the Hölder inequality imply that for all $\alpha \in \mathcal{A}$,

$$\mathbb{E}[|X(\tau_0)^T(\alpha - \alpha_0)|^3|Q] \leq 6\left\{\mathbb{E}[\{X(\tau_0)^T(\alpha - \alpha_0)\}^2|Q]\right\}^{3/2},$$

which provides a sufficient condition for condition (iii). On the other hand, our condition (iii) can hold more generally since (5.1) and (5.2) need to hold only locally around β_0 and θ_0, respectively.

Recall that in Assumption 3.4 (iii), we consider two cases: (i) $M_3^{-1} < \mathbb{E}[(X^T\delta_0)^2|Q = \tau] \leq M_3$ or (ii) $M_3^{-1}|\delta_0|_2^2 < \mathbb{E}[(X^T\delta_0)^2|Q = \tau] \leq M_3|\delta_0|_2^2$ uniformly in $\tau \in \mathcal{T}_0$ for some $M_3 > 0$. For the second case, we need to strengthen Assumption 5.1 (ii) slightly in the following way.

Assumption 5.2. *There are constants $C_3, \widetilde{\epsilon} > 0$ and a neighborhood $\mathcal{T}_0$ of τ_0 such that for all x in the support of X, for all q in $\mathcal{T}_0$, and for all ν such that $|\nu| \leq \widetilde{\epsilon}$,*

$$f_{Y|X,Q}(x(\tau_0)^T\alpha_0 + \nu x^T\delta_0|x, q) \geq C_3 > 0. \tag{5.3}$$

Assumption 5.2 is useful to verify Assumption 3.3 (iv) by insuring that τ_0 is well identified and the margin condition for τ_0 holds, even if $\inf_{\tau\in\mathcal{T}_0}\mathbb{E}[(X^T\delta_0)^2|Q = \tau]$ diverges to infinity. The following lemma verifies all the imposed regularity conditions on the loss function in Sections 3 and 4 for the quantile regression model.

Lemma 5.1. *(i) Let Assumption 5.1 hold. Moreover, for some $M_3 > 0$, we have that either $M_3^{-1} < \mathbb{E}[(X^T\delta_0)^2|Q = \tau] \leq M_3$ is satisfied uniformly in $\tau \in \mathcal{T}_0$ or Assumption 5.2 holds. Then Assumption 3.3 holds.*

(ii) In addition to Assumption 5.1, let Assumptions 3.1 (i), (iv), 3.2 (i), and 3.4 (i) hold. Suppose that α_0 is in the interior of $\mathcal{A}$. Then Assumption 4.1 holds with $M_n = Cs^{1/2}$ for some constant $C > 0$.

5.2 Logistic regression with a change point

Consider a binary outcome $Y \in \{0,1\}$, whose distribution depends on a high-dimensional regressor X with a possible change point:

$$P(Y=1|X,Q) = g(X^T\beta_0 + X^T\delta_0 1\{Q>\tau_0\}) = g(X(\tau_0)^T\alpha_0),$$

where $g(t) \equiv \exp(t)/[1+\exp(t)]$. Such a model belongs to the more general GLM family, but has independent interest in classifications and binary choice applications.

The loss function is given by the negative log-likelihood:

$$\rho(Y, X(\tau)^T\alpha) = -[Y \log g(X(\tau)^T\alpha) + (1-Y)\log(1-g(X(\tau)^T\alpha))].$$

It follows that

$$m_j(\tau,\alpha) = -E\left\{\left[\frac{g(X(\tau_0)^T\alpha_0)}{g(X(\tau)^T\alpha)} - \frac{1-g(X(\tau_0)^T\alpha_0)}{1-g(X(\tau)^T\alpha)}\right] g'(X_j(\tau)^T\alpha)X_j(\tau)\right\}$$

where $g'(t) = g(t)(1-g(t))$ denotes the derivative of $g(t)$. Immediately, $m_j(\tau_0,\alpha_0)=0$ for $j=1,...,2p$.

We make the following assumptions for logistic regression models.

Assumption 5.3. *There are $r>0$ and $\epsilon>0$ such that $\epsilon < g(X^T\beta), g(X^T\theta), g(X(\tau)^T\alpha) < 1-\epsilon$ almost surely for all $\beta \in \mathcal{B}(\beta_0,r)$, $\theta \in \mathcal{G}(\theta_0,r)$, $\alpha = (\beta^T, \theta^T - \beta^T)^T$ and $\tau \in (\tau_0 - r, \tau_0 + r)$.*

In particular, Assumption 5.3 requires that $g(X(\tau)^T\alpha)$ be bounded away from both zero and one in an ℓ_1 neighborhood of $\alpha_0 = (\beta_0^T, \theta_0^T - \beta_0^T)^T$. This assumption is restrictive but standard in that intuitively, we should have observations for both $Y=1$ and $Y=0$ almost everywhere within the support of (X,Q). In a similar way, van de Geer (2008) also mentions that her margin condition holds in the logistic regression if $\varepsilon \le \mathbb{E}(Y|X=x) \le 1-\varepsilon$ holds almost surely for some $\varepsilon > 0$. The following lemma verifies all the imposed regularity conditions on the loss function in Sections 3 and 4 for the logistic regression model.

Lemma 5.2. *(i) Let Assumption 5.3 hold. Then Assumption 3.3 holds.*

(ii) In addition to Assumption 5.3, let Assumptions 3.1 (i), (iv), 3.2 (i), and 3.4 (i) hold. Suppose that α_0 is in the interior of $\mathcal{A}$. Then Assumption 4.1 holds with $M_n = Cs^{1/2}$ for some constant $C>0$.

6 Monte Carlo Experiments

In this section we provide the results of some Monte Carlo simulation studies. We consider two simulation designs whose loss functions are convex: the median regression and the logistic regression. The data are generated respectively from the following two models:

$$Y_i = X_i'\beta_0 + 1\{Q_i < \tau_0\}X_i'\delta_0 + \varepsilon_{i1} \tag{6.1}$$

$$Y_i = 1\{X_i'\beta_0 + 1\{Q_i < \tau_0\}X_i'\delta_0 + \varepsilon_{i2} > 0\}, \tag{6.2}$$

where ε_{i1} is generated from the standard normal distribution, and ε_{i2} is generated from the logistic distribution. Similar to the existing simulation studies (e.g. Belloni and Chernozhukov (2011) and Wang (2013)), X_i is a p-dimensional vector generated from $N(0, \Sigma)$ whose columns are dependent by a covariance matrix $\Sigma_{ij} = (1/2)^{|i-j|}$. The threshold variable Q_i is a scalar generated from the uniform distribution on the interval of $(0, 1)$. The p-dimensional parameters β_0 and δ_0 are set to $\beta_0 = (0.5, 0, 0.5, 0, \ldots, 0)$ and $\delta = (0, 1, 1, 0, \ldots, 0)$ in case of (6.1) and $\beta_0 = (1.5, 0, 1.5, 0, \ldots, 0)$ and $\delta = (0, 3, 3, 0, \ldots, 0)$ in case of (6.2). The threshold parameter τ_0 is set to 0.5. The sample size is set to 400. We consider several different sizes of X_i, so that we set $p = 50, 100, 200$, and 400. Notice that the total number of regressors is $2p$ in each design. The range of τ is set to $\mathcal{T} = [0.15, 0.85]$. We conduct 1,000 replications of each design.

We estimate the model by the standard algorithm with slight modifications. In these simulation studies, we use the R package 'quantreg' for (6.1) and 'glmnet' for (6.2). For each design, we estimate the model using the above algorithms for each grid point of τ spanning over 71 equi-spaced points on $\mathbb{T}$. Next, we choose $\widehat{\tau}$ and corresponding $\widehat{\alpha}(\widehat{\tau})$ that minimize the objective function. Thus, we just need an additional loop over the grid points of τ. We conduct our simulation studies over several different tuning parameter values, whose results are quite similar. To save some space we report here simulation results from $\lambda_n = 0.03$ and $\mu_n = \log(p) \times \lambda_n$ for the quantile regression and from $\lambda_n = 0.03$ and $\mu_n = 0.5 \times \log(p) \times \lambda_n$ for the logistic regression.[1] Finally, We set $a = 3.7$ in the second step SCAD estimator following the convention in the literature (e.g. Fan and Li (2001) and Loh and Wainwright (2013)).

Tables 1–2 summarize these simulation results. In addition to the proposed estimator, we also report two oracle estimation results to evaluate the performance. Oracle 1 knows the true non-zero regressors $J(\alpha_0)$ and the threshold parameter τ_0 while Oracle 2 only knows $J(\alpha_0)$ and estimates τ as well as α. In the tables, we report the following statistics: the mean and median excess risks; the average number of non-zero parameters; the probability

[1] All tables across different tuning parameter values are available from the authors upon request.

Table 1: Median Regression

Design	Excess Risk		$\mathbb{E}[J(\widehat{\alpha})]$	$P\{J(\alpha_0) \subset J(\widehat{\alpha})\}$	$\mathbb{E}\,\lvert\widehat{\alpha}-\alpha_0\rvert_1$	$\mathbb{E}\,\lvert\widehat{\tau}-\tau_0\rvert_1$
	Mean	Median	$\left(\mathbb{E}[J(\widehat{\beta})]/\mathbb{E}[J(\widehat{\delta})]\right)$	$(\beta_{0,1}/\beta_{0,3}/\delta_{0,2}/\delta_{0,3})$	(on $J(\alpha_0)/J^c(\alpha_0)$)	
Oracle 1	0.003	0.002	NA	NA	0.318 (0.318 / NA)	NA
Oracle 2	0.004	0.003	NA	NA	0.319 (0.319 / NA)	0.003
$p = 50$	0.006	0.005	4.87 (2.8 / 2.1)	1 (1 / 1 / 1 / 1)	0.406 (0.381 / 0.025)	0.004
$p = 100$	0.008	0.006	4.58 (2.6 / 2.0)	0.99 (1 / 0.99 / 1 / 1)	0.465 (0.450 / 0.015)	0.005
$p = 200$	0.010	0.009	4.30 (2.3 / 2.0)	0.98 (1 / 0.98 / 1 / 1)	0.524 (0.517 / 0.007)	0.005
$p = 400$	0.014	0.012	4.09 (2.1 / 2.0)	0.96 (1 / 0.96 / 1 / 1)	0.619 (0.616 / 0.003)	0.005

Note: Oracle 1 knows both $J(\alpha_0)$ and τ_0 and Oracle 2 knows only $J(\alpha_0)$. For other designs, the tuning parameters are set to $\lambda_n = 0.03$ and $\mu_n = \log(p) \times \lambda_n$. The number of observations is set to $n = 400$. Expectations ($\mathbb{E}$) and the probability (P) are calculated by the average of 1,000 iterations in each design.

Table 2: Logistic Regression

Design	Excess Risk		$\mathbb{E}[J(\widehat{\alpha})]$	$P\{J(\alpha_0) \subset J(\widehat{\alpha})\}$	$\mathbb{E}\,\lvert\widehat{\alpha}-\alpha_0\rvert_1$	$\mathbb{E}\,\lvert\widehat{\tau}-\tau_0\rvert_1$
	Mean	Median	$\left(\mathbb{E}[J(\widehat{\beta})]/\mathbb{E}[J(\widehat{\delta})]\right)$	$(\beta_{0,1}/\beta_{0,3}/\delta_{0,2}/\delta_{0,3})$	(on $J(\alpha_0)/J^c(\alpha_0)$)	
Oracle 1	0.006	0.004	NA	NA	1.680 (1.680 / NA)	NA
Oracle 2	0.012	0.008	NA	NA	1.769 (1.769 / NA)	0.017
$p = 50$	0.017	0.012	4.40 (2.4 / 2.0)	0.95 (1 / 1 / 0.99 / 0.96)	1.957 (1.829 / 0.128)	0.026
$p = 100$	0.019	0.014	4.28 (2.3 / 2.0)	0.92 (1 / 1 / 1 / 0.93)	2.096 (1.991 / 0.105)	0.032
$p = 200$	0.021	0.014	4.22 (2.3 / 1.9)	0.9 (1 / 1 / 0.99 / 0.91)	2.256 (2.147 / 0.109)	0.042
$p = 400$	0.024	0.019	4.11 (2.2 / 1.9)	0.84 (1 / 1 / 0.96 / 0.88)	2.466 (2.363 / 0.103)	0.053

Note: Oracle 1 knows both $J(\alpha_0)$ and τ_0 and Oracle 2 knows only $J(\alpha_0)$. For other designs, the tuning parameters are set to $\lambda_n = 0.03$ and $\mu_n = 0.5 \times \log(p) \times \lambda_n$. The number of observations is set to $n = 400$. Expectations ($\mathbb{E}$) and the probability (P) are calculated by the average of 1,000 iterations in each design.

of containing true non-zero parameters; and ℓ_1 errors of $\widehat{\alpha}$ and $\widehat{\tau}$, respectively. We also report in the parentheses those statistics on the subset of the parameters.

Overall, the results are satisfactory and provide finite sample evidence for the theoretical results we develop in the previous sections. First, in Table 1 of the median regression model, the excess risk in both measures are small and very close to that of the oracle models. Furthermore, as we can see from the fourth and the fifth columns ($\mathbb{E}[J(\widehat{\alpha})]$ and $P\{J(\alpha_0) \subset J(\widehat{\alpha})\}$), the sparse model structure is well-captured and it seldomly misses the true non-zero parameters. The ℓ_1 errors of $\widehat{\alpha}$ and $\widehat{\tau}$ are also reasonably close to the oracle models. In Table 2 of the logistic regression model, the overall patterns of the results are very similar to those of the median regression although the size of errors is slightly larger. In both designs, the estimator works well even when the dimension of X_i is very large as in case of $2p = 800$. In sum, the proposed estimation procedure works well in finite samples and confirms the theoretical results developed earlier.

A Proofs for Section 3

Throughout the proof, we define

$$\nu_n(\alpha, \tau) \equiv \frac{1}{n} \sum_{i=1}^{n} \left[\rho\left(Y_i, X_i(\tau)^T \alpha\right) - \mathbb{E}\rho\left(Y, X(\tau)^T \alpha\right) \right].$$

Without loss of generality let $\nu_n(\alpha_J, \tau) = n^{-1} \sum_{i=1}^{n} \left[\rho\left(Y_i, X_{iJ}(\tau)^T \alpha_J\right) - \mathbb{E}\rho\left(Y, X_J(\tau)^T \alpha_J\right) \right]$. Also define $D(\tau) = \mathrm{diag}(D_j(\tau) : j \leq 2p)$ and then $D_0 = D(\tau_0)$ and $\widehat{D} = D(\widehat{\tau})$.

For the positive constant K_1 in Assumption 3.1 (i), define

$$c_{np} \equiv \sqrt{\frac{2\log(4np)}{n}} + \frac{K_1 \log(4np)}{n}.$$

Let $\lceil x \rceil$ denote the smallest integer greater than or equal to a real number x. The following lemma bounds $\nu_n(\alpha, \tau)$.

Lemma A.1. *For any positive sequences m_{1n} and m_{2n}, and any $\widetilde{\delta} \in (0,1)$, there are constants L_1, L_2 and $L_3 > 0$ such that for $a_n = L_1 c_{np} \widetilde{\delta}^{-1}$, $b_n = L_2 c_{np} \lceil \log_2(m_{2n}/m_{1n}) \rceil \widetilde{\delta}^{-1}$, and $c_n = L_3 n^{-1/2} \widetilde{\delta}^{-1}$,*

$$\mathbb{P}\left\{ \sup_{\tau \in \mathcal{T}} \sup_{|\alpha - \alpha_0|_1 \leq m_{1n}} |\nu_n(\alpha, \tau) - \nu_n(\alpha_0, \tau)| \geq a_n m_{1n} \right\} \leq \widetilde{\delta}, \tag{A.1}$$

$$\mathbb{P}\left\{\sup_{\tau\in\mathcal{T}}\sup_{m_{1n}\le|\alpha-\alpha_0|_1\le m_{2n}}\frac{|\nu_n(\alpha,\tau)-\nu_n(\alpha_0,\tau)|}{|\alpha-\alpha_0|_1}\ge b_n\right\}\le\widetilde{\delta}, \tag{A.2}$$

and for any $\eta>0$ *and* $\mathcal{T}_\eta=\{\tau\in\mathcal{T}:|\tau-\tau_0|\le\eta\}$,

$$\mathbb{P}\left\{\sup_{\tau\in\mathcal{T}_\eta}|\nu_n(\alpha_0,\tau)-\nu_n(\alpha_0,\tau_0)|\ge c_n|\delta_0|_2\sqrt{\eta}\right\}\le\widetilde{\delta}. \tag{A.3}$$

Proof of (A.1): Let $\epsilon_1,...,\epsilon_n$ denote a Rademacher sequence, independent of $\{Y_i,X_i,Q_i\}_{i\le n}$. By the symmetrization theorem (see, for example, Theorem 14.3 of Bühlmann and van de Geer (2011)) and then by the contraction theorem (see, for example, Theorem 14.4 of Bühlmann and van de Geer (2011)),

$$\begin{aligned}
&\mathbb{E}\left(\sup_{\tau\in\mathcal{T}}\sup_{|\alpha-\alpha_0|_1\le m_{1n}}|\nu_n(\alpha,\tau)-\nu_n(\alpha_0,\tau)|\right)\\
&\le 2\mathbb{E}\left(\sup_{\tau\in\mathcal{T}}\sup_{|\alpha-\alpha_0|_1\le m_{1n}}\left|\frac{1}{n}\sum_{i=1}^n\epsilon_i\left[\rho\left(Y_i,X_i(\tau)^T\alpha\right)-\rho\left(Y_i,X_i(\tau)^T\alpha_0\right)\right]\right|\right)\\
&\le 4L\mathbb{E}\left(\sup_{\tau\in\mathcal{T}}\sup_{|\alpha-\alpha_0|_1\le m_{1n}}\left|\frac{1}{n}\sum_{i=1}^n\epsilon_iX_i(\tau)^T(\alpha-\alpha_0)\right|\right).
\end{aligned}$$

Note that

$$\begin{aligned}
&\sup_{\tau\in\mathcal{T}}\sup_{|\alpha-\alpha_0|_1\le m_{1n}}\left|\frac{1}{n}\sum_{i=1}^n\epsilon_iX_i(\tau)^T(\alpha-\alpha_0)\right|\\
&=\sup_{\tau\in\mathcal{T}}\sup_{|\alpha-\alpha_0|_1\le m_{1n}}\left|\sum_{j=1}^{2p}(\alpha_j-\alpha_{0j})\frac{1}{n}\sum_{i=1}^n\epsilon_iX_{ij}(\tau)\right|\\
&\le\sup_{|\alpha-\alpha_0|_1\le m_{1n}}\sum_{j=1}^{2p}|\alpha_j-\alpha_{0j}|\sup_{\tau\in\mathcal{T}}\max_{j\le 2p}\left|\frac{1}{n}\sum_{i=1}^n\epsilon_iX_{ij}(\tau)\right|\\
&\le m_{1n}\sup_{\tau\in\mathcal{T}}\max_{j\le 2p}\left|\frac{1}{n}\sum_{i=1}^n\epsilon_iX_{ij}(\tau)\right|.
\end{aligned} \tag{A.4}$$

For all $\tilde{L} > K_1$,

$$\begin{aligned}
\mathbb{E}\left(\sup_{\tau\in\mathcal{T}}\max_{j\le 2p}\left|\sum_{i=1}^{n}\epsilon_i X_{ij}(\tau)\right|\right) &\le_{(1)} \tilde{L}\log\mathbb{E}\left[\exp\left(\tilde{L}^{-1}\sup_{\tau\in\mathcal{T}}\max_{j\le 2p}\left|\sum_{i=1}^{n}\epsilon_i X_{ij}(\tau)\right|\right)\right] \\
&\le_{(2)} \tilde{L}\log\mathbb{E}\left[\exp\left(\tilde{L}^{-1}\max_{\tau\in\{Q_1,\ldots,Q_n\}}\max_{j\le 2p}\left|\sum_{i=1}^{n}\epsilon_i X_{ij}(\tau)\right|\right)\right] \\
&\le_{(3)} \tilde{L}\log\left[4np\exp\left(\frac{n}{2(\tilde{L}^2-\tilde{L}K_1)}\right)\right],
\end{aligned}$$

where inequality (1) follows from Jensen's inequality, inequality (2) comes from the fact that $X_{ij}(\tau)$ is a step function with jump points on $\mathcal{T}\cap\{Q_1,\ldots,Q_n\}$, and inequality (3) is by Bernstein's inequality for the exponential moment of an average (see, for example, Lemma 14.8 of Bühlmann and van de Geer (2011)), combined with the simple inequalities that $\exp(|x|)\le\exp(x)+\exp(-x)$ and that $\exp(\max_{1\le j\le J}x_j)\le\sum_{j=1}^{J}\exp(x_j)$. Then it follows that

$$\mathbb{E}\left(\sup_{\tau\in\mathcal{T}}\max_{j\le 2p}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i X_{ij}(\tau)\right|\right)\le\frac{\tilde{L}\log(4np)}{n}+\frac{1}{2(\tilde{L}-K_1)}=c_{np}, \tag{A.5}$$

where the last equality follows by taking $\tilde{L}=K_1+\sqrt{n/[2\log(4np)]}$. Thus, by Markov's inequality,

$$\mathbb{P}\left\{\sup_{\tau\in\mathcal{T}}\sup_{|\alpha-\alpha_0|_1\le m_{1n}}|\nu_n(\alpha,\tau)-\nu_n(\alpha_0,\tau)|>a_n m_{1n}\right\}\le(a_n m_{1n})^{-1}4Lm_{1n}c_{np}=\widetilde{\delta},$$

where the last equality follows by setting $L_1=4L$.

Proof of (A.2): Recall that $\epsilon_1,\ldots,\epsilon_n$ is a Rademacher sequence, independent of $\{Y_i,X_i,Q_i\}_{i\le n}$.

Note that

$$
\begin{aligned}
&\mathbb{E}\left(\sup_{\tau\in\mathcal{T}}\sup_{m_{1n}\le|\alpha-\alpha_0|_1\le m_{2n}}\frac{|\nu_n(\alpha,\tau)-\nu_n(\alpha_0,\tau)|}{|\alpha-\alpha_0|_1}\right)\\
&\le_{(1)} 2\mathbb{E}\left(\sup_{\tau\in\mathcal{T}}\sup_{m_{1n}\le|\alpha-\alpha_0|_1\le m_{2n}}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\frac{\rho\left(Y_i,X_i(\tau)^T\alpha\right)-\rho\left(Y_i,X_i(\tau)^T\alpha_0\right)}{|\alpha-\alpha_0|_1}\right|\right)\\
&\le_{(2)} 2\sum_{j=1}^{k}\mathbb{E}\left(\sup_{\tau\in\mathcal{T}}\sup_{2^{j-1}m_{1n}\le|\alpha-\alpha_0|_1\le 2^{j}m_{1n}}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\frac{\rho\left(Y_i,X_i(\tau)^T\alpha\right)-\rho\left(Y_i,X_i(\tau)^T\alpha_0\right)}{2^{j-1}m_{1n}}\right|\right)\\
&\le_{(3)} 4L\sum_{j=1}^{k}\mathbb{E}\left(\sup_{\tau\in\mathcal{T}}\sup_{2^{j-1}m_{1n}\le|\alpha-\alpha_0|_1\le 2^{j}m_{1n}}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\frac{X_i(\tau)^T(\alpha-\alpha_0)}{2^{j-1}m_{1n}}\right|\right),
\end{aligned}
$$

where inequality (1) is by the symmetrization theorem, inequality (2) holds for some $k\equiv\lceil\log_2(m_{2n}/m_{1n})\rceil$, and inequality (3) follows from the contraction theorem.

Next, identical arguments showing (A.4) yield

$$
\sup_{2^{j-1}m_{1n}\le|\alpha-\alpha_0|_1\le 2^{j}m_{1n}}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\frac{X_i(\tau)^T(\alpha-\alpha_0)}{2^{j-1}m_{1n}}\right|\le 2\max_{j\le 2p}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i X_{ij}(\tau)\right|
$$

uniformly in $\tau\in\mathcal{T}$. Then, as in the proof of (A.1), by Bernstein's and Markov's inequalities,

$$
\mathbb{P}\left\{\sup_{\tau\in\mathcal{T}}\sup_{m_{1n}\le|\alpha-\alpha_0|_1\le m_{2n}}\frac{|\nu_n(\alpha,\tau)-\nu_n(\alpha_0,\tau)|}{|\alpha-\alpha_0|_1}>b_n\right\}\le b_n^{-1}8Lkc_{np}=\widetilde{\delta},
$$

where the last equality follows by setting $L_2=8L$.

Proof of (A.3): As above, by the symmetrization and contraction theorems, we have that

$$
\begin{aligned}
&\mathbb{E}\left(\sup_{\tau\in\mathcal{T}_\eta}|\nu_n(\alpha_0,\tau)-\nu_n(\alpha_0,\tau_0)|\right)\\
\le\ & 2\mathbb{E}\left(\sup_{\tau\in\mathcal{T}_\eta}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\left[\rho\left(Y_i,X_i(\tau)^T\alpha_0\right)-\rho\left(Y_i,X_i(\tau_0)^T\alpha_0\right)\right|\right]\right)\\
\le\ & 4L\mathbb{E}\left(\sup_{\tau\in\mathcal{T}_\eta}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i X_i^T\delta_0\left(1\{Q_i>\tau\}-1\{Q_i>\tau_0\}\right)\right|\right)\\
\le\ & \frac{4LC_1(M|\delta_0|_2^2K_2\eta)^{1/2}}{\sqrt{n}}
\end{aligned}
$$

for some constant $C_1<\infty$, where the last inequality is due to Theorem 2.14.1 of van der

Vaart and Wellner (1996) with M in Assumption 3.4 (i) and K_2 in Assumption 3.2 (i). Specifically, we apply the second inequality of this theorem to the class $\mathcal{F} = \{f(\epsilon, X, Q, \tau) = \epsilon X^T \delta_0 (1\{Q > \tau\} - 1\{Q > \tau_0\}), \tau \in \mathcal{T}_\eta\}$. Note that $\mathcal{F}$ is a Vapnik-Cervonenkis class, which has a uniformly bounded entropy integral and thus $J(1, \mathcal{F})$ in their theorem is bounded, and that the L_2 norm of the envelope $|\epsilon_i X_i^T \delta_0| 1\{|Q_i - \tau_0| < \eta\}$ is proportional to the square root of the length of $\mathcal{T}_\eta$:

$$(E|\epsilon_i X_i^T \delta_0|^2 1\{|Q_i - \tau_0| < \eta\})^{1/2} \leq (2M|\delta_0|_2^2 K_2 \eta)^{1/2}.$$

This implies the last inequality with C_1 being $\sqrt{2}$ times the entropy integral of the class $\mathcal{F}$. Then, by Markov's inequality, we obtain (A.3) with $L_3 = 4LC_1(MK_2)^{1/2}$.

A.1 Proof of Theorem 3.1

It follows from the definition of $(\widehat{\alpha}, \widehat{\tau})$ in (2.3) that

$$\frac{1}{n}\sum_{i=1}^{n} \rho(Y_i, X_i(\widehat{\tau})^T \widehat{\alpha}) + \lambda_n |\widehat{D}\widehat{\alpha}|_1 \leq \frac{1}{n}\sum_{i=1}^{n} \rho(Y_i, X_i(\tau_0)^T \alpha_0) + \lambda_n |D_0 \alpha_0|_1.$$

From this, we obtain the following inequality

$$\begin{aligned} R(\widehat{\alpha}, \widehat{\tau}) &\leq [\nu_n(\alpha_0, \tau_0) - \nu_n(\widehat{\alpha}, \widehat{\tau})] + \lambda_n |D_0\alpha_0|_1 - \lambda_n |\widehat{D}\widehat{\alpha}|_1 \\ &= [\nu_n(\alpha_0, \widehat{\tau}) - \nu_n(\widehat{\alpha}, \widehat{\tau})] + [\nu_n(\alpha_0, \tau_0) - \nu_n(\alpha_0, \widehat{\tau})] \\ &\quad + \lambda_n \left(|D_0\alpha_0|_1 - |\widehat{D}\widehat{\alpha}|_1 \right). \end{aligned} \tag{A.6}$$

Note that the second component $[\nu_n(\alpha_0, \tau_0) - \nu_n(\alpha_0, \widehat{\tau})] = o_P\left[(s/n)^{1/2} \log n\right]$ due to (A.3) of Lemma A.1 with taking $\mathcal{T}_\eta = \mathcal{T}$ by choosing some sufficiently large $\eta > 0$. Thus, we focus on the other two terms in the following discussion. We consider two cases respectively: $|\widehat{\alpha} - \alpha_0|_1 \leq |\alpha_0|_1$ and $|\widehat{\alpha} - \alpha_0|_1 > |\alpha_0|_1$.

Suppose that $|\widehat{\alpha} - \alpha_0|_1 \leq |\alpha_0|_1$. Then, $\left|\widehat{D}\widehat{\alpha}\right|_1 \leq \left|\widehat{D}\left(\widehat{\alpha} - \alpha_0\right)\right|_1 + \left|\widehat{D}\alpha_0\right|_1 \leq 2\bar{D}\left|\alpha_0\right|_1$, and

$$\left|\lambda_n \left(|D_0\alpha_0|_1 - |\widehat{D}\widehat{\alpha}|_1 \right)\right| \leq 3\lambda_n \bar{D} \left|\alpha_0\right|_1.$$

Apply (A.1) in Lemma A.1 with $m_{1n} = |\alpha_0|_1$ to obtain

$$|\nu(\alpha_0, \widehat{\tau}) - \nu_n(\widehat{\alpha}, \widehat{\tau})| \leq a_n \left|\alpha_0\right|_1 \leq \lambda_n \left|\alpha_0\right|_1,$$

with probability approaching one (w.p.a.1), where the last inequality follows from the fact that $a_n \ll \lambda_n$ since $\lambda_n = C\omega_n$ for some constant $C > 0$ with ω_n defined in (3.5). Thus, the theorem follows in this case.

Now assume that $|\widehat{\alpha} - \alpha_0|_1 > |\alpha_0|_1$. In this case, apply (A.2) of Lemma A.1 with $m_{1n} = |\alpha_0|_1$ and $m_{2n} = 2M_1 p$, where M_1 is defined in Assumption 3.1(ii), to obtain

$$\frac{|\nu(\alpha_0, \widehat{\tau}) - \nu_n(\widehat{\alpha}, \widehat{\tau})|}{|\widehat{\alpha} - \alpha_0|_1} \leq b_n \quad \text{w.p.a.1.}$$

Since $b_n \ll \underline{D}\lambda_n$, we have, with probability approaching one,

$$|\nu(\alpha_0, \widehat{\tau}) - \nu_n(\widehat{\alpha}, \widehat{\tau})| \leq \lambda_n \underline{D} \, |\widehat{\alpha} - \alpha_0|_1 \leq \lambda_n \left| \widehat{D} \left(\widehat{\alpha} - \alpha_0 \right) \right|_1 .$$

Therefore,

$$\begin{aligned} R(\widehat{\alpha}, \widehat{\tau}) + o_P\left(n^{-1/2} \log n\right) &\leq \lambda_n \left(|D_0 \alpha_0|_1 - |\widehat{D}\widehat{\alpha}|_1 \right) + \lambda_n \left| \widehat{D} \left(\widehat{\alpha} - \alpha_0 \right) \right|_1 \\ &\leq \lambda_n \left(|D_0 \alpha_0|_1 - |\widehat{D}\widehat{\alpha}_J|_1 \right) + \lambda_n \left| \widehat{D} \left(\widehat{\alpha} - \alpha_0 \right)_J \right|_1 , \end{aligned}$$

where the last inequality follows from the fact that $\widehat{\alpha} - \alpha_0 = \widehat{\alpha}_{J^C} + \left(\widehat{\alpha} - \alpha_0 \right)_J$. Thus, the theorem follows in this case as well.

A.2 Proof of Theorem 3.2

Recall from (3.7) that for all $\alpha = (\beta^T, \delta^T)^T \in \mathbb{R}^{2p}$ and $\theta = \beta + \delta$, the excess risk has the following decomposition: when $\tau > \tau_0$,

$$\begin{aligned} R\left(\alpha, \tau\right) = \; &\mathbb{E}\left(\left[\rho\left(Y, X^T \beta\right) - \rho\left(Y, X^T \beta_0\right) \right] 1\left\{ Q \leq \tau_0 \right\} \right) \\ &+ \mathbb{E}\left(\left[\rho\left(Y, X^T \theta\right) - \rho\left(Y, X^T \theta_0\right) \right] 1\left\{ Q > \tau \right\} \right) \\ &+ \mathbb{E}\left(\left[\rho\left(Y, X^T \beta\right) - \rho\left(Y, X^T \theta_0\right) \right] 1\left\{ \tau_0 < Q \leq \tau \right\} \right) . \end{aligned} \tag{A.7}$$

We split the proof into four steps.

Step 1: All the three terms on the right hand side (RHS) of (A.7) are nonnegative. As a consequence, all the three terms on the RHS of (A.7) are bounded by $R(\alpha, \tau)$.

Proof of Step 1. Step 1 is implied by the condition that $\mathbb{E}[\rho(Y, X(\tau_0)^T \alpha) - \rho(Y, X(\tau_0)^T \alpha_0) | Q] \geq 0$ a.s. for all $\alpha \in \mathcal{A}$. To see this, the first two terms are nonnegative by simply multiplying $\mathbb{E}[\rho(Y, X(\tau_0)^T \alpha) - \rho(Y, X(\tau_0)^T \alpha_0) | Q] \geq 0$ with $1\{Q \leq \tau_0\}$ and $1\{Q > \tau\}$ respectively. To show that the third term is nonnegative for all $\beta \in \mathbb{R}^p$ and $\tau > \tau_0$, set $\alpha = (\beta/2, \beta/2)$ in the inequality $1\{\tau_0 < Q \le \tau\}\mathbb{E}[\rho(Y, X(\tau_0)^T\alpha) - \rho(Y, X(\tau_0)^T\alpha_0)|Q] \ge 0$. Then we have that

$$1\{\tau_0 < Q \le \tau\}\mathbb{E}[\rho(Y, X^T(\beta/2 + \beta/2)) - \rho(Y, X^T\theta_0)|Q] \ge 0,$$

which yields the nonnegativeness of the third term. ∎

Step 2: Let $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$. Prove:

$$\mathbb{E}\left[|X^T(\beta - \beta_0)|1\{Q \le \tau_0\}\right] \le \frac{1}{\eta^* r^*}R(\alpha, \tau) \vee \left[\frac{1}{\eta^*}R(\alpha, \tau)\right]^{1/2}.$$

Proof of Step 2. Recall that

$$r_1(\eta) \equiv \sup_r \Big\{r : \mathbb{E}\left(\left[\rho\left(Y, X^T\beta\right) - \rho\left(Y, X^T\beta_0\right)\right] 1\left\{Q \le \tau_0\right\}\right) \\ \ge \eta\mathbb{E}[(X^T(\beta - \beta_0))^2 1\{Q \le \tau_0\}] \text{ for all } \beta \in \mathcal{B}(\beta_0, r)\Big\}.$$

For notational simplicity, write

$$\mathbb{E}[(X^T(\beta - \beta_0))^2 1\{Q \le \tau_0\}] \equiv \|\beta - \beta_0\|_q^2,$$

and

$$F(\delta) \equiv \mathbb{E}\left(\left[\rho\left(Y, X^T(\beta_0 + \delta)\right) - \rho\left(Y, X^T\beta_0\right)\right] 1\left\{Q \le \tau_0\right\}\right).$$

Note that $F(\beta - \beta_0) = \mathbb{E}\left(\left[\rho\left(Y, X^T\beta\right) - \rho\left(Y, X^T\beta_0\right)\right] 1\left\{Q \le \tau_0\right\}\right)$, and $\beta \in \mathcal{B}(\beta_0, r)$ if and only if $\|\beta - \beta_0\|_q \le r$.

For any β, if $\|\beta - \beta_0\|_q \le r_1(\eta^*)$, then by the definition of $r_1(\eta^*)$, we have:

$$F(\beta - \beta_0) \ge \eta^*\mathbb{E}[(X^T(\beta - \beta_0))^2 1\{Q \le \tau_0\}].$$

If $\|\beta - \beta_0\|_q > r_1(\eta^*)$, let $t = r_1(\eta^*)\|\beta - \beta_0\|_q^{-1} \in (0, 1)$. Since $F(.)$ is convex, and $F(0) = 0$, we have $F(\beta - \beta_0) \ge t^{-1}F(t(\beta - \beta_0))$. Moreover, define

$$\check{\beta} = \beta_0 + r_1(\eta^*)\frac{\beta - \beta_0}{\|\beta - \beta_0\|_q},$$

then $\|\check{\beta} - \beta_0\|_q = r_1(\eta^*)$ and $t(\beta - \beta_0) = \check{\beta} - \beta_0$. Hence still by the definition of $r_1(\eta^*)$,

$$F(\beta - \beta_0) \ge \frac{1}{t}F(\check{\beta} - \beta_0) \ge \frac{\eta^*}{t}\mathbb{E}[(X^T(\check{\beta} - \beta_0))^2 1\{Q \le \tau_0\}] = \eta^* r_1(\eta^*)\|\beta - \beta_0\|_q.$$

Therefore, by Assumption 3.3 (iii), and Step 1,

$$\begin{aligned}R(\alpha,\tau) &\geq \mathbb{E}\left(\left[\rho\left(Y, X^T\beta\right) - \rho\left(Y, X^T\beta_0\right)\right] 1\left\{Q \leq \tau_0\right\}\right)\\ &\geq \eta^*\mathbb{E}[(X^T(\beta-\beta_0))^2 1\{Q\leq\tau_0\}] \wedge \eta^* r^* \{\mathbb{E}[(X^T(\beta-\beta_0))^2 1\{Q\leq\tau_0\}]\}^{1/2}\\ &\geq \eta^*\left(\mathbb{E}\left[|X^T(\beta-\beta_0)|1\{Q\leq\tau_0\}\right]\right)^2 \wedge \eta^* r^* \mathbb{E}\left[|X^T(\beta-\beta_0)|1\{Q\leq\tau_0\}\right],\end{aligned}$$

where the last inequality follows from Jensen's inequality. ∎

Step 3: For any $\epsilon' > 0$, there is an $\varepsilon > 0$ such that for all $\tau > \tau_0$, and $\alpha \in \mathbb{R}^{2p}$, $R(\alpha,\tau) < \varepsilon$ implies $|\tau - \tau_0| < \epsilon'$.

Proof of Step 3. We first prove that, for any $\epsilon' > 0$, there is $\varepsilon > 0$ such that for all $\tau > \tau_0$, and $\alpha \in \mathbb{R}^{2p}$, $R(\alpha,\tau) < \varepsilon$ implies that $\tau < \tau_0 + \epsilon'$.

Suppose that $R(\alpha,\tau) < \varepsilon$. Applying the triangle inequality, for all β and $\tau > \tau_0$,

$$\begin{aligned}&\mathbb{E}\left[\left(\rho\left(Y, X^T\beta_0\right) - \rho\left(Y, X^T\theta_0\right)\right) 1\left\{\tau_0 < Q \leq \tau\right\}\right]\\ &\leq \left|\mathbb{E}\left[\left(\rho\left(Y, X^T\beta\right) - \rho\left(Y, X^T\theta_0\right)\right) 1\left\{\tau_0 < Q \leq \tau\right\}\right]\right|\\ &+ \left|\mathbb{E}\left[\left(\rho\left(Y, X^T\beta\right) - \rho\left(Y, X^T\beta_0\right)\right) 1\left\{\tau_0 < Q \leq \tau\right\}\right]\right|.\end{aligned} \tag{A.8}$$

First, note that the first term on the RHS of (A.8) is the third term on the RHS of (A.7), hence is bounded by $R(\alpha,\tau) < \varepsilon$.

We now consider the second term on the RHS of (A.8). Assumption 3.5 (i) implies, with $C_1^T = C_1^{-1}(1 - C_1) > 0$ and $C_2^T = C_2^{-1}(1 - C_2) > 0$, for all $\beta \in \mathbb{R}^p$,

$$C_2^T\mathbb{E}\left[|X^T\beta| 1\left\{Q > \tau_0\right\}\right] \leq \mathbb{E}\left[|X^T\beta| 1\left\{Q \leq \tau_0\right\}\right] \leq C_1^T\mathbb{E}\left[|X^T\beta| 1\left\{Q > \tau_0\right\}\right]. \tag{A.9}$$

It follows from the Lipschitz condition (Assumption 3.3 (i)), Step 2, and (A.9) that

$$\begin{aligned}\left|\mathbb{E}\left[\left(\rho\left(Y, X^T\beta\right) - \rho\left(Y, X^T\beta_0\right)\right) 1\left\{\tau_0 < Q \leq \tau\right\}\right]\right| &\leq L\mathbb{E}\left[\left|X^T\left(\beta-\beta_0\right)\right| 1\left\{\tau_0 < Q \leq \tau\right\}\right]\\ &\leq L\mathbb{E}\left[\left|X^T\left(\beta-\beta_0\right)\right| 1\left\{\tau_0 < Q\right\}\right]\\ &\leq LC_2^{T^{-1}}\mathbb{E}\left[\left|X^T\left(\beta-\beta_0\right)\right| 1\left\{Q \leq \tau_0\right\}\right]\\ &\leq LC_2^{T^{-1}}\left\{\varepsilon/(\eta^* r^*) \vee \sqrt{\varepsilon/\eta^*}\right\}\\ &\equiv C(\varepsilon).\end{aligned}$$

Thus, we have shown that (A.8) is bounded by $C(\varepsilon) + \varepsilon$.

For any $\epsilon' > 0$, it follows from Assumptions 3.3 (ii), 3.3 (iv) and 3.4 (ii) (see also Remark

3.2) that there is a $c > 0$ such that if $\tau > \tau_0 + \epsilon'$,

$$\begin{aligned} c\mathbb{P}\left(\tau_0 < Q \leq \tau_0 + \epsilon'\right) &\leq c\mathbb{P}\left(\tau_0 < Q \leq \tau\right) \\ &\leq \mathbb{E}\left[\left(\rho\left(Y, X^T\beta_0\right) - \rho\left(Y, X^T\theta_0\right)\right) 1\left\{\tau_0 < Q \leq \tau\right\}\right] \\ &\leq C(\varepsilon) + \varepsilon. \end{aligned}$$

Since $\varepsilon \mapsto C(\varepsilon) + \varepsilon$ converges to zero as ε converges to zero, for a given $\epsilon' > 0$ choose a sufficient small $\varepsilon > 0$ such that $C(\varepsilon) + \varepsilon < c\mathbb{P}(\tau_0 < Q \leq \tau_0 + \epsilon')$, so that the above inequality cannot hold. Hence we infer that for this ε, when $R(\alpha, \tau) < \varepsilon$, we must have $\tau < \tau_0 + \epsilon'$.

By the same argument, if $\tau < \tau_0$, then we must have $\tau > \tau_0 - \epsilon'$. Hence, $R(\alpha, \tau) < \varepsilon$ implies $|\tau - \tau_0| < \epsilon'$. ∎

Step 4: $\widehat{\tau} \xrightarrow{p} \tau_0$.

Proof of Step 4. For the ε chosen in Step 3, consider the event $\{R(\widehat{\alpha}, \widehat{\tau}) < \varepsilon\}$, which occurs with probability approaching one due to Theorem 3.1. On this event, $|\widehat{\tau} - \tau_0| < \epsilon'$ by Step 3. Because ϵ' is taken arbitrarily, we have proved the consistency of $\widehat{\tau}$. ∎

A.3 Proof of Theorem 3.3

The proof consists of several steps. First, we prove that $\widehat{\beta}$ and $\widehat{\theta}$ are inside the neighborhoods of β_0 and θ_0, respectively. Second, we obtain an intermediate convergence rate for $\widehat{\tau}$ based on the consistency of the risk and $\widehat{\tau}$. Finally, we use the compatibility condition to obtain a tighter bound.

Step 1: For any $r > 0$, with probability approaching one (w.p.a.1), $\widehat{\beta} \in \mathcal{B}(\beta_0, r)$ and $\widehat{\theta} \in \mathcal{G}(\theta_0, r)$.

Proof of Step 1. Suppose that $\widehat{\tau} > \tau_0$. The proof of Step 2 in the proof of Theorem 3.2 implies that when $\tau > \tau_0$,

$$\mathbb{E}\left[(X^T(\beta - \beta_0))^2 1\{Q \leq \tau_0\}\right] \leq \frac{R(\alpha, \tau)^2}{(\eta^* r^*)^2} \vee \frac{R(\alpha, \tau)}{\eta^*}.$$

For any $r > 0$, note that $R(\widehat{\alpha}, \widehat{\tau}) = o_P(1)$ implies that the event $R(\widehat{\alpha}, \widehat{\tau}) < r^2$ holds w.p.a.1. Therefore, we have shown that $\widehat{\beta} \in \mathcal{B}(\beta_0, r)$.

We now show that $\widehat{\theta} \in \mathcal{G}(\theta_0, r)$. When $\tau > \tau_0$, we have that

$$
\begin{aligned}
R(\alpha,\tau) &\geq_{(1)} \mathbb{E}\left(\left[\rho\left(Y, X^T\theta\right) - \rho\left(Y, X^T\theta_0\right)\right] 1\left\{Q > \tau\right\}\right) \\
&= \mathbb{E}\left(\left[\rho\left(Y, X^T\theta\right) - \rho\left(Y, X^T\theta_0\right)\right] 1\left\{Q > \tau_0\right\}\right) \\
&- \mathbb{E}\left(\left[\rho\left(Y, X^T\theta\right) - \rho\left(Y, X^T\theta_0\right)\right] 1\left\{\tau \geq Q > \tau_0\right\}\right) \\
&\geq_{(2)} \eta^* \mathbb{E}\left[|X^T(\theta - \theta_0)|^2 1\{Q > \tau_0\}\right] \wedge \eta^* r^* \left(\mathbb{E}\left[|X^T(\theta - \theta_0)|^2 1\{Q > \tau_0\}\right]\right)^{1/2} \\
&\quad - \mathbb{E}\left(\left[\rho\left(Y, X^T\theta\right) - \rho\left(Y, X^T\theta_0\right)\right] 1\left\{\tau \geq Q > \tau_0\right\}\right),
\end{aligned}
$$

where (1) is from (3.7) and (2) can be proved using arguments similar to those used in the proof of Step 2 in the proof of Theorem 3.2. This implies that

$$
\mathbb{E}\left[(X^T(\theta - \theta_0))^2 1\{Q > \tau_0\}\right] \leq \frac{\tilde{R}(\alpha,\tau)^2}{(\eta^* r^*)^2} \vee \frac{\tilde{R}(\alpha,\tau)}{\eta^*}.
$$

where $\tilde{R}(\alpha,\tau) \equiv R(\alpha,\tau) + \mathbb{E}\left(\left[\rho\left(Y, X^T\theta\right) - \rho\left(Y, X^T\theta_0\right)\right] 1\left\{\tau \geq Q > \tau_0\right\}\right)$. Thus, it suffices to show that $\tilde{R}(\widehat{\alpha}, \widehat{\tau}) = o_P(1)$ in order to establish that $\widehat{\theta} \in \mathcal{G}(\theta_0, r)$. Note that for some constant $C > 0$,

$$
\begin{aligned}
&\mathbb{E}\left[(\rho(Y, X^T\theta) - \rho(Y, X^T\theta_0)) 1\{\tau_0 < Q \leq \tau\}\right] \\
&\leq_{(1)} L\mathbb{E}\left[|X^T(\theta - \theta_0)| 1\{\tau_0 < Q \leq \tau\}\right] \\
&\leq_{(2)} L|\theta - \theta_0|_1 \mathbb{E}\left[\max_{j \leq p} |\tilde{X}_j| 1\{\tau_0 < Q \leq \tau\}\right] + L|\theta - \theta_0|_1 \mathbb{E}\left[|Q| 1\{\tau_0 < Q \leq \tau\}\right] \\
&\leq_{(3)} L|\theta - \theta_0|_1 \mathbb{E}\left[\max_{j \leq p} |\tilde{X}_j| \sup_{\tilde{x}} \mathbb{P}(\tau_0 < Q \leq \tau | \tilde{X} = \tilde{x})\right] + L|\theta - \theta_0|_1 \mathbb{E}\left[|Q| 1\{\tau_0 < Q \leq \tau\}\right] \\
&\leq_{(4)} C(\tau - \tau_0)|\theta - \theta_0|_1 \mathbb{E}\left\{\left[\max_{j \leq p} |\tilde{X}_j|\right] + 1\right\},
\end{aligned}
$$

where (1) is by the Lipschitz continuity of $\rho(Y, \cdot)$, (2) is from the fact that $|X^T(\theta - \theta_0)| \leq |\theta - \theta_0|_1(\max_{j \leq p} |\tilde{X}_j| + |Q|)$, (3) is by taking the conditional probability, (4) is from Assumption 3.2 (iii).

By the expectation-form of the Bernstein inequality (Lemma 14.12 of Bühlmann and van de Geer (2011)), $\mathbb{E}[\max_{j \leq p} |X_j|] \leq K_1 \log(p+1) + \sqrt{2\log(p+1)}$. By (A.15), which will be shown below, $|\widehat{\theta} - \theta_0|_1 = O_P(s)$. Hence by (A.11), when $\widehat{\tau} > \tau_0$,

$$
|\widehat{\tau} - \tau_0| |\widehat{\theta} - \theta_0|_1 \mathbb{E}[\max_{j \leq p} |X_j|] = O_P(\lambda_n s^2 \log p) = o_P(1).
$$

Note that when $\widehat{\tau} > \tau_0$, the proofs of (A.15) and (A.11) do not require $\widehat{\theta} \in \mathcal{G}(\theta_0, r)$, so there is no problem of applying them here. This implies that $\tilde{R}(\widehat{\alpha},\widehat{\tau}) = o_P(1)$.

The same argument yields that w.p.a.1, $\widehat{\theta} \in \mathcal{G}(\theta_0, r)$ and $\widehat{\beta} \in \mathcal{B}(\beta_0, r)$ when $\widehat{\tau} \le \tau_0$; hence it is omitted to avoid repetitions. ∎

Step 2: Let $\bar{c}_0(\delta_0) \equiv c_0 \inf_{\tau \in \mathcal{T}_0} \mathbb{E}[(X^T\delta_0)^2|Q=\tau]$. Then $\bar{c}_0(\delta_0)\,|\widehat{\tau}-\tau_0| \le 4R\,(\widehat{\alpha},\widehat{\tau})$ w.p.a.1. As a result, $|\widehat{\tau}-\tau_0| = O_P\left[\lambda_n s/\bar{c}_0(\delta_0)\right]$.

Proof. For any $\tau_0 < \tau$ and $\tau \in \mathcal{T}_0$, and any $\beta \in \mathcal{B}(\beta_0, r)$, $\alpha = (\beta, \delta)$ with arbitrary δ, for some $L, M > 0$ which do not depend on β and τ,

$$
\begin{aligned}
&\left|\mathbb{E}\left(\rho\left(Y, X^T\beta\right) - \rho\left(Y, X^T\beta_0\right)\right) 1\left\{\tau_0 < Q \le \tau\right\}\right| \\
&\le_{(1)} L\mathbb{E}\left[\left|X^T\left(\beta-\beta_0\right)\right| 1\left\{\tau_0 < Q \le \tau\right\}\right] \\
&\le_{(2)} ML(\tau-\tau_0)\mathbb{E}\left[\left|X^T\left(\beta-\beta_0\right)\right| 1\left\{Q \le \tau_0\right\}\right] \\
&\le_{(3)} ML(\tau-\tau_0)\left\{\mathbb{E}\left[\left(X^T\left(\beta-\beta_0\right)\right)^2 1\left\{Q \le \tau_0\right\}\right]\right\}^{1/2} \\
&\le_{(4)} \left(ML(\tau-\tau_0)\right)^2/\left(4\eta^*\right) + \eta^*\mathbb{E}\left[\left(X^T\left(\beta-\beta_0\right)\right)^2 1\left\{Q \le \tau_0\right\}\right] \\
&\le_{(5)} \left(ML(\tau-\tau_0)\right)^2/\left(4\eta^*\right) + \mathbb{E}\left[\left(\rho\left(Y, X^T\beta\right) - \rho\left(Y, X^T\beta_0\right)\right) 1\left\{Q \le \tau_0\right\}\right] \\
&\le_{(6)} \left(ML(\tau-\tau_0)\right)^2/\left(4\eta^*\right) + R(\alpha,\tau),
\end{aligned}
$$

where (1) follows from the Lipschitz condition on the objective function (see Assumption 3.3 (i)), (2) is by Assumption 3.5 (ii), (3) is by Jensen's inequality, (4) follows from the fact that $uv \le v^2/(4c) + cu^2$ for any $c > 0$, (5) is from Assumption 3.3 (iii), and (6) is from Step 1 in the proof of Theorem 3.2.

In addition,

$$
\begin{aligned}
&\left|\mathbb{E}\left[\left(\rho\left(Y, X^T\beta\right) - \rho\left(Y, X^T\beta_0\right)\right) 1\left\{\tau_0 < Q \le \tau\right\}\right]\right| \\
&\ge_{(1)} \mathbb{E}\left[\left(\rho\left(Y, X^T\beta_0\right) - \rho\left(Y, X^T\theta_0\right)\right) 1\left\{\tau_0 < Q \le \tau\right\}\right] \\
&- \left|\mathbb{E}\left[\left(\rho\left(Y, X^T\beta\right) - \rho\left(Y, X^T\theta_0\right)\right) 1\left\{\tau_0 < Q \le \tau\right\}\right]\right| \\
&\ge_{(2)} \mathbb{E}\left[\left(\rho\left(Y, X^T\beta_0\right) - \rho\left(Y, X^T\theta_0\right)\right) 1\left\{\tau_0 < Q \le \tau\right\}\right] - R(\alpha,\tau) \\
&\ge_{(3)} c_0\left\{\inf_{\tau \in \mathcal{T}_0} \mathbb{E}[(X^T\delta_0)^2|Q=\tau]\right\}(\tau-\tau_0) - R(\alpha,\tau),
\end{aligned}
$$

where (1) is by the triangular inequality, (2) is from (3.7), and (3) is by Assumption 3.3 (iv). Therefore, we have established that there exists a constant $\tilde{C} > 0$, independent of (α, τ), such that

$$
\bar{c}_0(\delta_0)(\tau-\tau_0) \le \tilde{C}(\tau-\tau_0)^2 + 2R(\alpha,\tau). \tag{A.10}
$$

Note that when $0 < (\tau - \tau_0) < \bar{c}_0(\delta_0)(2\tilde{C})^{-1}$, (A.10) implies that

$$\bar{c}_0(\delta_0)(\tau - \tau_0) \leq \frac{\bar{c}_0(\delta_0)}{2}(\tau - \tau_0) + 2R(\alpha, \tau),$$

which in turn implies that $\tau - \tau_0 \leq \frac{4}{\bar{c}_0(\delta_0)} R(\alpha, \tau)$. By the same argument, when $-\bar{c}_0(\delta_0)(2\tilde{C})^{-1} < (\tau - \tau_0) \leq 0$, we have $\tau_0 - \tau \leq \frac{4}{\bar{c}_0(\delta_0)} R(\alpha, \tau)$ for $\alpha = (\beta, \delta)$, with any $\theta \in \mathcal{G}(\theta_0, r)$ and arbitrary β.

Hence when $\widehat{\tau} > \tau_0$, on the event $\widehat{\beta} \in \mathcal{B}(\beta_0, r)$, and $\widehat{\tau} - \tau_0 < \bar{c}_0(\delta_0)(2\tilde{C})^{-1}$, we have

$$\widehat{\tau} - \tau_0 \leq \frac{4}{\bar{c}_0(\delta_0)} R(\widehat{\alpha}, \widehat{\tau}). \tag{A.11}$$

When $\widehat{\tau} \leq \tau_0$, on the event $\widehat{\theta} \in \mathcal{G}(\theta_0, r)$, and $\tau_0 - \widehat{\tau} < \bar{c}_0(\delta_0)(2\tilde{C})^{-1}$, we have $\tau_0 - \widehat{\tau} \leq \frac{4}{\bar{c}_0(\delta_0)} R(\widehat{\alpha}, \widehat{\tau})$. Hence due to Step 1 and the consistency of $\widehat{\tau}$, we have, w.p.a.1,

$$|\widehat{\tau} - \tau_0| \leq \frac{4}{\bar{c}_0(\delta_0)} R\left(\widehat{\alpha}, \widehat{\tau}\right). \tag{A.12}$$

This also implies $|\widehat{\tau} - \tau_0| = O_P\left[\lambda_n s / \bar{c}_0(\delta_0)\right]$ in view of the proof of Theorem 3.1. ∎

Step 3: Define $\nu_{1n}(\tau) \equiv \nu_n(\alpha_0, \tau) - \nu_n(\alpha_0, \tau_0)$ and $c_\alpha \equiv \lambda_n \left(|D_0 \alpha_0|_1 - \left|\widehat{D}\alpha_0\right|_1 \right) + |\nu_{1n}(\widehat{\tau})|$. Then w.p.a.1,

$$R\left(\widehat{\alpha}, \widehat{\tau}\right) + \frac{1}{2}\lambda_n \left|\widehat{D}\left(\widehat{\alpha} - \alpha_0\right)\right|_1 \leq c_\alpha + 2\lambda_n \left|\widehat{D}\left(\widehat{\alpha} - \alpha_0\right)_J\right|_1. \tag{A.13}$$

Proof. Recall the following basic inequality in (A.6):

$$R(\widehat{\alpha}, \widehat{\tau}) \leq [\nu_n(\alpha_0, \widehat{\tau}) - \nu_n(\widehat{\alpha}, \widehat{\tau})] - \nu_{1n}(\widehat{\tau}) + \lambda_n \left(|D_0 \alpha_0|_1 - |\widehat{D}\widehat{\alpha}|_1 \right). \tag{A.14}$$

Now applying Lemma A.1 (in particular, (A.1)) to $[\nu_n(\alpha_0, \widehat{\tau}) - \nu_n(\widehat{\alpha}, \widehat{\tau})]$ with a_n and b_n replaced by $a_n/2$ and $b_n/2$, we can rewrite the basic inequality in (A.14) by

$$\lambda_n |D_0 \alpha_0|_1 \geq R\left(\widehat{\alpha}, \widehat{\tau}\right) + \lambda_n \left|\widehat{D}\widehat{\alpha}\right|_1 - \frac{1}{2}\lambda_n \left|\widehat{D}\left(\widehat{\alpha} - \alpha_0\right)\right|_1 - |\nu_{1n}(\widehat{\tau})| \quad \text{w.p.a.1.}$$

Now adding $\lambda_n \left|\widehat{D}\left(\widehat{\alpha} - \alpha_0\right)\right|_1$ on both sides of the inequality above and using the fact that

$|\alpha_{0j}|_1 - |\widehat{\alpha}_j|_1 + |(\widehat{\alpha}_j - \alpha_{0j})|_1 = 0$ for $j \notin J$, we have that w.p.a.1,

$$\begin{aligned}
&\lambda_n\left(|D_0\alpha_0|_1 - \left|\widehat{D}\alpha_0\right|_1\right) + |\nu_{1n}(\widehat{\tau})| + 2\lambda_n\left|\widehat{D}(\widehat{\alpha}-\alpha_0)_J\right|_1 \\
&\geq R(\widehat{\alpha},\widehat{\tau}) + \frac{1}{2}\lambda_n\left|\widehat{D}(\widehat{\alpha}-\alpha_0)\right|_1.
\end{aligned}$$

Therefore, we have proved Step 3. ∎

We prove the remaining part of the steps by considering two cases: (i) $\lambda_n\left|\widehat{D}(\widehat{\alpha}-\alpha_0)_J\right|_1 \leq c_\alpha$; (ii) $\lambda_n\left|\widehat{D}(\widehat{\alpha}-\alpha_0)_J\right|_1 > c_\alpha$. We first consider Case (ii).

Step 4: Suppose that $\lambda_n\left|\widehat{D}(\widehat{\alpha}-\alpha_0)_J\right|_1 > c_\alpha$. Then

$$|\widehat{\tau} - \tau_0| = O_P\left[\lambda_n^2 s/\bar{c}_0(\delta_0)\right] \quad \text{and} \quad |\widehat{\alpha} - \alpha_0| = O_P(\lambda_n s).$$

Proof. By $\lambda_n\left|\widehat{D}(\widehat{\alpha}-\alpha_0)_J\right|_1 > c_\alpha$ and the basic inequality (A.13) in Step 3,

$$6\left|\widehat{D}(\widehat{\alpha}-\alpha_0)_J\right|_1 \geq \left|\widehat{D}(\widehat{\alpha}-\alpha_0)\right|_1 = \left|\widehat{D}(\widehat{\alpha}-\alpha_0)_J\right| + \left|\widehat{D}(\widehat{\alpha}-\alpha_0)_{J^c}\right|, \tag{A.15}$$

which enables us to apply the compatibility condition in Assumption 3.6.

Recall that $\|Z\|_2 = (EZ^2)^{1/2}$ for a random variable Z. Note that for $s = |J(\alpha_0)|_0$,

$$\begin{aligned}
&R(\widehat{\alpha},\widehat{\tau}) + \frac{1}{2}\lambda_n\left|\widehat{D}(\widehat{\alpha}-\alpha_0)\right|_1 \\
&\leq_{(1)} 3\lambda_n\left|\widehat{D}(\widehat{\alpha}-\alpha_0)_J\right|_1 \\
&\leq_{(2)} 3\lambda_n\bar{D}\left\|X(\widehat{\tau})^T(\widehat{\alpha}-\alpha_0)\right\|_2\sqrt{s}/\phi \\
&\leq_{(3)} \frac{9\lambda_n^2\bar{D}^2 s}{2\tilde{c}\phi^2} + \frac{\tilde{c}}{2}\left\|X(\widehat{\tau})^T(\widehat{\alpha}-\alpha_0)\right\|_2^2,
\end{aligned} \tag{A.16}$$

where (1) is from the basic inequality (A.13) in Step 3, (2) is by the compatibility condition (Assumption 3.6), and (3) is from the inequality that $uv \leq v^2/(2\tilde{c}) + \tilde{c}u^2/2$ for any $\tilde{c} > 0$.

We will show below in Step 5 that there is a constant $C_0 > 0$ such that

$$\left\|X(\widehat{\tau})^T(\widehat{\alpha}-\alpha_0)\right\|_2^2 \leq C_0 R(\widehat{\alpha},\widehat{\tau}) + C_0\bar{c}_0(\delta_0)|\widehat{\tau}-\tau_0|, \text{ w.p.a.1.} \tag{A.17}$$

Recall that by (A.12), $\bar{c}_0(\delta_0)|\widehat{\tau}-\tau_0| \leq 4R(\widehat{\alpha},\widehat{\tau})$. Hence, (A.16) with $\tilde{c} = (C_0 + \frac{4C_0}{\bar{c}_0(\delta_0)})^{-1}$

implies that

$$R\left(\widehat{\alpha},\widehat{\tau}\right)+\lambda_n\left|\widehat{D}\left(\widehat{\alpha}-\alpha_0\right)\right|_1\leq\frac{9\lambda_n^2\bar{D}^2s}{\tilde{c}\phi^2}. \tag{A.18}$$

By (A.18) and (A.12), $|\widehat{\tau}-\tau_0|=O_P\left[\lambda_n^2 s/\bar{c}_0(\delta_0)\right]$. Also, by (A.18), $|\widehat{\alpha}-\alpha_0|=O_P\left(\lambda_n s\right)$ since $D(\widehat{\tau})\geq\underline{D}$ w.p.a.1 by Assumption 3.1 (iii). ∎

Step 5: There is a constant $C_0>0$ such that $\left\|X(\widehat{\tau})^T(\widehat{\alpha}-\alpha_0)\right\|_2^2\leq C_0R(\widehat{\alpha},\widehat{\tau})+C_0\bar{c}_0(\delta_0)|\widehat{\tau}-\tau_0|$, w.p.a.1.

Proof. Note that

$$\begin{aligned}\left\|X(\tau)^T(\alpha-\alpha_0)\right\|_2^2&\leq 2\left\|X(\tau)^T\alpha-X(\tau_0)^T\alpha\right\|_2^2\\&\quad+4\left\|X(\tau_0)^T\alpha-X(\tau_0)^T\alpha_0\right\|_2^2+4\left\|X(\tau_0)^T\alpha_0-X(\tau)^T\alpha_0\right\|_2^2.\end{aligned} \tag{A.19}$$

We bound the three terms on the right hand side of (A.19). When $\tau>\tau_0$, there is a constant $C_1>0$ such that

$$\begin{aligned}&\left\|X(\tau)^T\alpha-X(\tau_0)^T\alpha\right\|_2^2\\&=\mathbb{E}\left[(X^T\delta)^2 1\{\tau_0\leq Q<\tau\}\right]\\&=\int_{\tau_0}^{\tau}\mathbb{E}\left[(X^T\delta)^2\middle|Q=t\right]dF_Q(t)\\&\leq 2\int_{\tau_0}^{\tau}\mathbb{E}\left[(X^T\delta_0)^2\middle|Q=t\right]dF_Q(t)+2\int_{\tau_0}^{\tau}\mathbb{E}\left[(X^T(\delta-\delta_0))^2\middle|Q=t\right]dF_Q(t)\\&\leq C_1\bar{c}_0(\delta_0)(\tau-\tau_0),\end{aligned}$$

where the last inequality is by Assumptions 3.1 (i), 3.2 (iii), 3.4 (iii), and 3.5 (ii).

Similarly, $\left\|X(\tau_0)^T\alpha_0-X(\tau)^T\alpha_0\right\|_2^2=\mathbb{E}\left[(X^T\delta_0)^2 1\{\tau_0\leq Q<\tau\}\right]\leq C_1\bar{c}_0(\delta_0)(\tau-\tau_0)$. Hence, the first and third terms of the right hand side of of (A.19) are bounded by $6C_1\bar{c}_0(\delta_0)(\tau-\tau_0)$.

To bound the second term, note that there exists a constant $C_2 > 0$ such that

$$
\begin{aligned}
&\left\|X(\tau_0)^T\alpha - X(\tau_0)^T\alpha_0\right\|_2^2 \\
&=_{(1)} \mathbb{E}\left[(X^T(\theta-\theta_0))^2 1\{Q>\tau_0\}\right] + \mathbb{E}\left[(X^T(\beta-\beta_0))^2 1\{Q\le\tau_0\}\right] \\
&\le_{(2)} (\eta^*)^{-1}\mathbb{E}\left[\left(\rho\left(Y,X^T\theta\right)-\rho\left(Y,X^T\theta_0\right)\right)1\{Q>\tau_0\}\right] \\
&+ (\eta^*)^{-1}\mathbb{E}\left[\left(\rho\left(Y,X^T\beta\right)-\rho\left(Y,X^T\beta_0\right)\right)1\{Q\le\tau_0\}\right] \\
&\le_{(3)} (\eta^*)^{-1}R(\alpha,\tau) + (\eta^*)^{-1}\mathbb{E}\left[\left(\rho\left(Y,X^T\theta\right)-\rho\left(Y,X^T\theta_0\right)\right)1\{\tau_0<Q\le\tau\}\right] \\
&\le_{(4)} (\eta^*)^{-1}R(\alpha,\tau) + (\eta^*)^{-1}L\mathbb{E}\left[|X^T(\theta-\theta_0)|1\{\tau_0<Q\le\tau\}\right] \\
&=_{(5)} (\eta^*)^{-1}R(\alpha,\tau) + (\eta^*)^{-1}L\int_{\tau_0}^{\tau}\mathbb{E}\left[|X^T(\theta-\theta_0)|\big|Q=t\right]dF_Q(t) \\
&\le_{(6)} (\eta^*)^{-1}R(\alpha,\tau) + C_3(\tau-\tau_0),
\end{aligned}
$$

where (1) is simply an identity, (2) from Assumption 3.3 (iii), (3) is due to (A.7): namely,

$$
\mathbb{E}\left[\left(\rho\left(Y,X^T\theta\right)-\rho\left(Y,X^T\theta_0\right)\right)1\{Q>\tau\}\right] + \mathbb{E}\left[\left(\rho\left(Y,X^T\beta\right)-\rho\left(Y,X^T\beta_0\right)\right)1\{Q\le\tau_0\}\right] \le R(\alpha,\tau),
$$

(4) is by the Lipschitz continuity of $\rho(Y,\cdot)$, (5) is by rewriting the expectation term, and (6) is by Assumptions 3.2 (i) and 3.5 (ii). Therefore, we have shown that $\left\|X(\tau)^T(\alpha-\alpha_0)\right\|_2^2 \le C_0R(\alpha,\tau) + C_0\bar{c}_0(\delta_0)(\tau-\tau_0)$ for some constant $C_0 > 0$. The case of $\tau \le \tau_0$ can be proved using the same argument. Hence, setting $\tau = \widehat{\tau}$, and $\alpha = \widehat{\alpha}$, we obtain the desired result. ∎

Step 6: We now consider Case (i). Suppose that $\lambda_n \left|\widehat{D}\left(\widehat{\alpha}-\alpha_0\right)_J\right| \le c_\alpha$. Then

$$
|\widehat{\tau}-\tau_0| = O_P\left[\lambda_n^2 s/\bar{c}_0(\delta_0)\right] \quad \text{and} \quad |\widehat{\alpha}-\alpha_0| = O_P\left(\lambda_n s\right).
$$

Proof. Recall that X_{ij} is the jth element of X_i, where $i \le n, j \le p$. By Assumption 3.1 (iv) and Step 2,

$$
\sup_{1\le j\le p}\frac{1}{n}\sum_{i=1}^n |X_{ij}|^2\left|1\left(Q_i<\widehat{\tau}\right)-1\left(Q_i<\tau_0\right)\right| = O_P\left[\lambda_n s/\bar{c}_0(\delta_0)\right].
$$

By the mean value theorem,

$$
\begin{aligned}
&\lambda_n \left| |D_0\alpha_0|_1 - \left|\widehat{D}\alpha_0\right|_1 \right| \\
&\leq \lambda_n \sum_{j=1}^{p} \left(\frac{2}{n} \sum_{i=1}^{n} |X_{ij} 1\{Q_i > \underline{\tau}\}|^2 \right)^{-1/2} \left|\delta_0^{(j)}\right| \frac{1}{n} \sum_{i=1}^{n} |X_{ij}|^2 \left|1\{Q_i > \widehat{\tau}\} - 1\{Q_i > \tau_0\}\right| \\
&= O_P\left[\lambda_n^2 s |J(\delta_0)|_0 / \bar{c}_0(\delta_0)\right]. \qquad \text{(A.20)}
\end{aligned}
$$

Here, recall that $\underline{\tau}$ is the left-end point of $\mathcal{T}$ and $|J(\delta_0)|_0$ is the dimension of nonzero elements of δ_0.

Due to Step 2 and (A.3) in Lemma A.1,

$$
|\nu_{1n}(\widehat{\tau})| = O_P\left[\frac{|\delta_0|_2}{\sqrt{\bar{c}_0(\delta_0)}} (\lambda_n s/n)^{1/2} \right]. \qquad \text{(A.21)}
$$

Thus, under Case (i), we have that, by (A.12), (A.13), (A.15), and (A.20),

$$
\begin{aligned}
\frac{\bar{c}_0(\delta_0)}{4} |\widehat{\tau} - \tau_0| &\leq \frac{\lambda_n}{2} \left|\widehat{D}(\widehat{\alpha} - \alpha_0)\right|_1 + R(\widehat{\alpha}, \widehat{\tau}) \\
&\leq 3\lambda_n \left(|D_0\alpha_0|_1 - \left|\widehat{D}\alpha_0\right|_1 \right) + 3 |\nu_{1n}(\widehat{\tau})| \qquad \text{(A.22)} \\
&= O_P\left(\lambda_n^2 s^2\right) + O_P\left[s^{1/2} (\lambda_n s/n)^{1/2} \right],
\end{aligned}
$$

where the last equality uses the fact that $|J(\delta_0)|_0/\bar{c}_0(\delta_0) = O(s)$ and $|\delta_0|_2/\sqrt{\bar{c}_0(\delta_0)} = O(s^{1/2})$ at most (both could be bounded in some cases).

Therefore, we now have an improved rate of convergence in probability for $\widehat{\tau}$ from $r_{n0,\tau} \equiv \lambda_n s$ to $r_{n1,\tau} \equiv [\lambda_n^2 s^2 + s^{1/2}(\lambda_n s/n)^{1/2}]$. Repeating the arguments identical to those to prove (A.20) and (A.21) yields that

$$
\lambda_n \left| |D_0\alpha_0|_1 - \left|\widehat{D}\alpha_0\right|_1 \right| = O_P\left[r_{n1,\tau}\lambda_n s\right] \quad \text{and} \quad |\nu_{1n}(\widehat{\tau})| = O_P\left[s^{1/2} (r_{n1,\tau}/n)^{1/2} \right].
$$

Plugging these improved rates into (A.22) gives

$$
\begin{aligned}
\bar{c}_0(\delta_0) |\widehat{\tau} - \tau_0| &= O_P\left(\lambda_n^3 s^3\right) + O_P\left[s^{1/2}(\lambda_n s)^{3/2}/n^{1/2}\right] + O_P\left(\lambda_n s^{3/2}/n^{1/2}\right) + O_P\left[s^{3/4}(\lambda_n s)^{1/4}/n^{3/4}\right] \\
&= O_P\left(\lambda_n^2 s^{3/2}\right) + O_P\left[s^{3/4}(\lambda_n s)^{1/4}/n^{3/4}\right] \\
&\equiv O_P(r_{n2,\tau}),
\end{aligned}
$$

where the second equality comes from the fact that the first three terms are $O_P\left(\lambda_n^2 s^{3/2}\right)$ since $\lambda_n s^{3/2} = o(1)$, $\lambda_n n/s \to \infty$, and $\lambda_n \sqrt{n} \to \infty$ in view of the assumption that $\lambda_n s^2 \log p = o(1)$.

Repeating the same arguments again with the further improved rate $r_{n2,\tau}$, we have that

$$|\widehat{\tau}-\tau_0| = O_P\left(\lambda_n^2 s^{5/4}\right) + O_P\left[s^{7/8}(\lambda_n s)^{1/8}/n^{7/8}\right] \equiv O_P(r_{n3,\tau}).$$

Thus, repeating the same arguments k times yields

$$\bar{c}_0(\delta_0)\,|\widehat{\tau}-\tau_0| = O_P\left(\lambda_n^2 s^{1+2^{-k}}\right) + O_P\left[s^{(2^k-1)/2^k}(\lambda_n s)^{1/2^k}/n^{(2^k-1)/2^k}\right] \equiv O_P(r_{nk,\tau}).$$

Then letting $k\to\infty$ gives the desired result that $\bar{c}_0(\delta_0)\,|\widehat{\tau}-\tau_0| = O_P\left(\lambda_n^2 s\right)$. Finally, the same iteration based on (A.22) gives $\left|\widehat{D}\left(\widehat{\alpha}-\alpha_0\right)\right| = o_P\left(\lambda_n s\right)$, which proves the desired result since $D(\widehat{\tau}) \geq \underline{D}$ w.p.a.1 by Assumption 3.1 (iii). ∎

A.4 Proof of Theorem 3.4

If $\delta_0 = 0$, τ_0 is non-identifiable. In this case, we decompose the excess risk in the following way:

$$\begin{aligned} R\left(\alpha,\tau\right) = {} & \mathbb{E}\left(\left[\rho\left(Y, X^T\beta\right) - \rho\left(Y, X^T\beta_0\right)\right] 1\left\{Q \leq \tau\right\}\right) \\ & + \mathbb{E}\left(\left[\rho\left(Y, X^T\theta\right) - \rho\left(Y, X^T\beta_0\right)\right] 1\left\{Q > \tau\right\}\right). \end{aligned} \tag{A.23}$$

We split the proof into three steps.

Step 1: For any $r>0$, we have that w.p.a.1, $\widehat{\beta} \in \tilde{\mathcal{B}}(\beta_0, r, \widehat{\tau})$ and $\widehat{\theta} \in \tilde{\mathcal{G}}(\beta_0, r, \widehat{\tau})$.

Proof of Step 1. As in the proof of Step 1 in the proof of Theorem 3.3, Assumption 3.7 (ii) implies that

$$\mathbb{E}\left[(X^T(\beta-\beta_0))^2 1\{Q \leq \tau\}\right] \leq \frac{R(\alpha,\tau)^2}{(\eta^* r^*)^2} \vee \frac{R(\alpha,\tau)}{\eta^*}.$$

For any $r>0$, note that $R(\widehat{\alpha},\widehat{\tau}) = o_P(1)$ implies that the event $R(\widehat{\alpha},\widehat{\tau}) < r^2$ holds w.p.a.1. Therefore, we have shown that $\widehat{\beta} \in \tilde{\mathcal{B}}(\beta_0, r, \widehat{\tau})$. The other case can be proved similarly. ∎

Step 2 : Suppose that $\delta_0 = 0$. Then

$$R\left(\widehat{\alpha},\widehat{\tau}\right) + \frac{1}{2}\lambda_n \left|\widehat{D}\left(\widehat{\alpha}-\alpha_0\right)\right|_1 \leq 2\lambda_n \left|\widehat{D}\left(\widehat{\alpha}-\alpha_0\right)_J\right|_1 \text{ w.p.a.1.} \tag{A.24}$$

Proof. The proof of this step is similar to that of Step 3 in the proof of Theorem 3.3. Since

$(\widehat{\alpha}, \widehat{\tau})$ minimizes the ℓ_1-penalized objective function in (2.3), we have that

$$\frac{1}{n}\sum_{i=1}^{n} \rho(Y_i, X_i(\widehat{\tau})^T\widehat{\alpha}) + \lambda_n|\widehat{D}\widehat{\alpha}|_1 \leq \frac{1}{n}\sum_{i=1}^{n} \rho(Y_i, X_i(\widehat{\tau})^T\alpha_0) + \lambda_n|\widehat{D}\alpha_0|_1. \tag{A.25}$$

When $\delta_0 = 0$, $\rho(Y, X(\widehat{\tau})^T\alpha_0) = \rho(Y, X(\tau_0)^T\alpha_0)$. Using this fact and (A.25), we obtain the following inequality

$$R(\widehat{\alpha}, \widehat{\tau}) \leq [\nu_n(\alpha_0, \widehat{\tau}) - \nu_n(\widehat{\alpha}, \widehat{\tau})] + \lambda_n|\widehat{D}\alpha_0|_1 - \lambda_n|\widehat{D}\widehat{\alpha}|_1. \tag{A.26}$$

As in Step 3 in the proof of Theorem 3.3, we apply Lemma A.1 (in particular, (A.1)) to $[\nu_n(\alpha_0, \widehat{\tau}) - \nu_n(\widehat{\alpha}, \widehat{\tau})]$ with a_n and b_n replaced by $a_n/2$ and $b_n/2$. Then we can rewrite the basic inequality in (A.26) by

$$\lambda_n \left|\widehat{D}\alpha_0\right|_1 \geq R(\widehat{\alpha}, \widehat{\tau}) + \lambda_n \left|\widehat{D}\widehat{\alpha}\right|_1 - \frac{1}{2}\lambda_n \left|\widehat{D}(\widehat{\alpha} - \alpha_0)\right|_1 \quad \text{w.p.a.1.}$$

Now adding $\lambda_n \left|\widehat{D}(\widehat{\alpha} - \alpha_0)\right|_1$ on both sides of the inequality above and using the fact that $|\alpha_{0j}|_1 - |\widehat{\alpha}_j|_1 + |(\widehat{\alpha}_j - \alpha_{0j})|_1 = 0$ for $j \notin J$, we have that w.p.a.1,

$$2\lambda_n \left|\widehat{D}(\widehat{\alpha} - \alpha_0)_J\right|_1 \geq R(\widehat{\alpha}, \widehat{\tau}) + \frac{1}{2}\lambda_n \left|\widehat{D}(\widehat{\alpha} - \alpha_0)\right|_1.$$

Therefore, we have obtained the desired result. ∎

Step 3 : Suppose that $\delta_0 = 0$. Then

$$R(\widehat{\alpha}, \widehat{\tau}) = O_P(\lambda_n^2 s) \quad \text{and} \quad |\widehat{\alpha} - \alpha_0| = O_P(\lambda_n s).$$

Proof. By Step 2,

$$4\left|\widehat{D}(\widehat{\alpha} - \alpha_0)_J\right|_1 \geq \left|\widehat{D}(\widehat{\alpha} - \alpha_0)\right|_1 = \left|\widehat{D}(\widehat{\alpha} - \alpha_0)_J\right| + \left|\widehat{D}(\widehat{\alpha} - \alpha_0)_{J^c}\right|, \tag{A.27}$$

which enables us to apply the compatibility condition in Assumption 3.8.

Recall that $\|Z\|_2 = (EZ^2)^{1/2}$ for a random variable Z. Note that for $s = |J(\alpha_0)|_0$,

$$
\begin{aligned}
& R(\widehat{\alpha}, \widehat{\tau}) + \frac{1}{2}\lambda_n \left|\widehat{D}(\widehat{\alpha} - \alpha_0)\right|_1 \\
& \leq_{(1)} 2\lambda_n \left|\widehat{D}(\widehat{\alpha} - \alpha_0)_J\right|_1 \\
& \leq_{(2)} 2\lambda_n \bar{D} \left\|X(\widehat{\tau})^T(\widehat{\alpha} - \alpha_0)\right\|_2 \sqrt{s}/\phi \\
& \leq_{(3)} \frac{4\lambda_n^2 \bar{D}^2 s}{2\tilde{c}\phi^2} + \frac{\tilde{c}}{2}\left\|X(\widehat{\tau})^T(\widehat{\alpha} - \alpha_0)\right\|_2^2,
\end{aligned}
\tag{A.28}
$$

where (1) is from the basic inequality (A.24) in Step 2, (2) is by the compatibility condition (Assumption 3.8), and (3) is from the inequality that $uv \leq v^2/(2\tilde{c}) + \tilde{c}u^2/2$ for any $\tilde{c} > 0$.

Note that

$$
\begin{aligned}
& \left\|X(\tau)^T\alpha - X(\tau)^T\alpha_0\right\|_2^2 \\
& =_{(1)} \mathbb{E}\left[(X^T(\theta - \beta_0))^2 1\{Q > \tau\}\right] + \mathbb{E}\left[(X^T(\beta - \beta_0))^2 1\{Q \leq \tau\}\right] \\
& \leq_{(2)} (\eta^*)^{-1}\mathbb{E}\left[\left(\rho\left(Y, X^T\theta\right) - \rho\left(Y, X^T\beta_0\right)\right) 1\{Q > \tau\}\right] \\
& + (\eta^*)^{-1}\mathbb{E}\left[\left(\rho\left(Y, X^T\beta\right) - \rho\left(Y, X^T\beta_0\right)\right) 1\{Q \leq \tau\}\right] \\
& \leq_{(3)} (\eta^*)^{-1} R(\alpha, \tau),
\end{aligned}
$$

where (1) is simply an identity, (2) from Assumption 3.7 (ii) , and (3) is due to (A.23). Hence, (A.28) with $\tilde{c} = \eta^*$ implies that

$$
R(\widehat{\alpha}, \widehat{\tau}) + \lambda_n \left|\widehat{D}(\widehat{\alpha} - \alpha_0)\right|_1 \leq \frac{4\lambda_n^2 \bar{D}^2 s}{\eta^* \phi^2}. \tag{A.29}
$$

Therefore, $R(\widehat{\alpha}, \widehat{\tau}) = O_P(\lambda_n^2 s)$. Also, $|\widehat{\alpha} - \alpha_0| = O_P(\lambda_n s)$ since $D(\widehat{\tau}) \geq \underline{D}$ w.p.a.1 by Assumption 3.1 (iii). ∎

B Proofs for Section 4

B.1 Proof of Theorem 4.1

We write α_J be a subvector of α whose components' indices are in $J(\alpha_0)$. Define $\bar{Q}_n(\alpha_J) \equiv \widetilde{S}_n((\alpha_J, 0))$, so that

$$
\bar{Q}_n(\alpha_J) = \frac{1}{n}\sum_{i=1}^{n} \rho(Y_i, X_{iJ}(\widehat{\tau})^T\alpha_J) + \mu_n \sum_{j \in J(\alpha_0)} w_j \widehat{D}_j |\alpha_j|.
$$

For notational simplicity, here we write $\widehat{D}_j \equiv D_j(\widehat{\tau})$.

Our proofs below go through for both the two cases: (i) $\delta_0 \neq 0$ and τ_0 is identifiable, and (ii) $\delta_0 = 0$ so τ_0 is not identifiable. The proofs of Theorems 4.1 and 4.2 are combined. Throughout the proofs, when τ_0 is identifiable, our argument is conditional on

$$\widehat{\tau} \in \mathcal{T}_n = \left\{ |\tau - \tau_0| \leq \omega_n^2 s \cdot \log n \right\}, \tag{B.1}$$

whose probability goes to 1 due to Theorem 3.3. On the other hand, when τ_0 is not identifiable, $\delta_0 = 0$, $\widehat{\tau}$ obtained in the first-step estimation can be any value in $\mathcal{T}$.

We first prove the following two lemmas. Define

$$\bar{\alpha}_J \equiv \operatorname*{argmin}_{\alpha_J} \bar{Q}_n(\alpha_J). \tag{B.2}$$

Lemma B.1. *Suppose that $s^4(\log p)^3(\log n)^3 + sM_n^4(\log p)^6(\log n)^6 = o(n)$ and $\widehat{\tau} \in \mathcal{T}_n$ if $\delta_0 \neq 0$; suppose that $s^4 \log s = o(n)$ and $\widehat{\tau}$ is any value in $\mathcal{T}$ if $\delta_0 = 0$. Then*

$$|\bar{\alpha}_J - \alpha_{0J}|_2 = O_P\left(\sqrt{\frac{s \log s}{n}}\right).$$

Proof of Lemma B.1. Let $k_n = \sqrt{\frac{s \log s}{n}}$. We first prove that for any $\epsilon > 0$, there is $C_\epsilon > 0$, with probability at least $1 - \epsilon$,

$$\inf_{|\alpha_J - \alpha_{0J}|_2 = C_\epsilon k_n} \bar{Q}_n(\alpha_J) > \bar{Q}_n(\alpha_{0J}) \tag{B.3}$$

Once this is proved, then by the continuity of $\bar{Q}_n$, there is a local minimizer of $\bar{Q}_n(\alpha_J)$ inside $B(\alpha_{0J}, C_\epsilon k_n) \equiv \{\alpha_J \in \mathbb{R}^s : |\alpha_{0J} - \alpha_J|_2 \leq C_\epsilon k_n\}$. Due to the convexity of $\bar{Q}_n$, such a local minimizer is also global. We now prove (B.3).

Write

$$l_J(\alpha_J) = \frac{1}{n} \sum_{i=1}^{n} \rho(Y_i, X_{iJ}(\widehat{\tau})^T \alpha_J), \quad L_J(\alpha_J, \tau) = \mathbb{E}[\rho(Y, X_J(\tau)^T \alpha_J)].$$

Then for all $|\alpha_J-\alpha_{0J}|_2=C_\epsilon k_n$,

$$\begin{aligned}
&\bar{Q}_n(\alpha_J)-\bar{Q}_n(\alpha_{0J})\\
=\;& l_J(\alpha_J)-l_J(\alpha_{0J})+\sum_{j\in J(\alpha_0)} w_j\mu_n\widehat{D}_j(|\alpha_j|-|\alpha_{0j}|)\\
\geq\;& \underbrace{L_J(\alpha_J,\widehat{\tau})-L_J(\alpha_{0J},\widehat{\tau})}_{(1)}-\underbrace{\sup_{|\alpha_J-\alpha_{0J}|_2\leq C_\delta k_n}|\nu_n(\alpha_J,\widehat{\tau})-\nu_n(\alpha_{0J},\widehat{\tau})|}_{(2)}+\underbrace{\sum_{j\in J(\alpha_0)}\mu_n\widehat{D}_j w_j(|\alpha_j|-|\alpha_{0j}|)}_{(3)}.
\end{aligned}$$

To analyze (1), note that $|\alpha_J-\alpha_{0J}|_2=C_\epsilon k_n$ and $m_J(\tau_0,\alpha_0)=0$ and when $\delta_0=0$, $m_J(\tau,\alpha_{0J})$ is free of τ. Then there is $c_3>0$,

$$\begin{aligned}
&L_J(\alpha_J,\widehat{\tau})-L_J(\alpha_{0J},\widehat{\tau})\\
\geq\;& m_J(\tau_0,\alpha_{0J})^T(\alpha_J-\alpha_{0J})+(\alpha_J-\alpha_{0J})^T\frac{\partial^2\mathbb{E}[\rho(Y,X_J(\widehat{\tau})^T\alpha_{0J})]}{\partial\alpha_J\partial\alpha_J^T}(\alpha_J-\alpha_{0J})\\
&-|m_J(\tau_0,\alpha_{0J})-m_J(\widehat{\tau},\alpha_{0J})|_2|\alpha_J-\alpha_{0J}|_2-c_3|\alpha_{0J}-\alpha_J|_1^3\\
\geq\;& \lambda_{\min}\left(\frac{\partial^2\mathbb{E}[\rho(Y,X_J(\widehat{\tau})^T\alpha_{0J})]}{\partial\alpha_J\partial\alpha_J^T}\right)|\alpha_J-\alpha_{0J}|_2^2\\
&-(|m_J(\tau_0,\alpha_{0J})-m_J(\widehat{\tau},\alpha_{0J})|_2)|\alpha_J-\alpha_{0J}|_2-c_3s^{3/2}|\alpha_{0J}-\alpha_J|_2^3\\
\geq\;& c_1C_\epsilon^2k_n^2-(|m_J(\tau_0,\alpha_{0J})-m_J(\widehat{\tau},\alpha_{0J})|_2)C_\epsilon k_n-c_3s^{3/2}C_\delta^3k_n^3\\
\geq\;& C_\epsilon k_n(c_1C_\epsilon k_n-M_n\omega_n^2s\cdot\log n-c_3s^{3/2}C_\epsilon^2k_n^2)\geq c_1C_\delta^2k_n^2/3,
\end{aligned}$$

where the last inequality follows from $M_n\omega_n^2s\cdot\log n<1/3c_1C_\epsilon k_n$ and $c_3s^{3/2}C_\epsilon^2k_n^2)<1/3c_1C_\epsilon k_n$. These follow from the condition $s^4\log s+sM_n^4(\log p)^6(\log n)^6=o(n)$

To analyze (2), by the symmetrization theorem and the contraction theorem (see, for example, Theorems 14.3 and 14.4 of Bühlmann and van de Geer (2011)), there is a Rademacher sequence $\epsilon_1,...,\epsilon_n$ independent of $\{Y_i,X_i,Q_i\}_{i\leq n}$ such that (note that when $\delta_0=0$, $\alpha_J=\beta_J$,

$$\nu_n(\alpha_J,\tau)\equiv\frac{1}{n}\sum_{i=1}^n\left[\rho\left(Y_i,X_{J(\beta_0)i}^T\beta_J\right)-\mathbb{E}\rho\left(Y,X_{J(\beta_0)}^T\beta_J\right)\right],$$

which is free of τ.)

$$
\begin{aligned}
V_n &= \mathbb{E}\left(\sup_{\tau\in\mathcal{T}_n}\sup_{|\alpha_J-\alpha_{0J}|_2\le C_\epsilon k_n}|\nu_n(\alpha_J,\tau)-\nu_n(\alpha_{0J},\tau)|\right)\\
&\le 2\mathbb{E}\left(\sup_{\tau\in\mathcal{T}_n}\sup_{|\alpha_J-\alpha_{0J}|_2\le C_\epsilon k_n}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i[\rho(Y_i,X_{iJ}(\tau)^T\alpha_J)-\rho(Y_i,X_{iJ}(\tau)^T\alpha_{0J})]\right|\right)\\
&\le 4L\mathbb{E}\left(\sup_{\tau\in\mathcal{T}_n}\sup_{|\alpha_J-\alpha_{0J}|_2\le C_\epsilon k_n}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i(X_{iJ}(\tau)^T(\alpha_J-\alpha_{0J}))\right|\right),
\end{aligned}
$$

which is bounded by the sum of the following two terms, $V_{1n}+V_{2n}$, due to the triangle inequality and the fact that $|\alpha_J-\alpha_{0J}|_1\le|\alpha_J-\alpha_{0J}|_2\sqrt{s}$: when $\delta_0\ne 0$ and τ_0 is identifiable,

$$
\begin{aligned}
V_{1n} &= 4L\mathbb{E}\left(\sup_{\tau\in\mathcal{T}_n}\sup_{|\alpha_J-\alpha_{0J}|_1\le C_\epsilon k_n\sqrt{s}}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i(X_{iJ}(\tau)-X_{iJ}(\tau_0))^T(\alpha_J-\alpha_{0J})\right|\right)\\
&\le 4L\mathbb{E}\left(\sup_{\tau\in\mathcal{T}_n}\sup_{|\delta_J-\delta_{0J}|_1\le C_\epsilon k_n\sqrt{s}}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i X_{iJ(\delta_0)}^T(1\{Q_i>\tau\}-1\{Q_i>\tau_0\})(\delta_J-\delta_{0J})\right|\right)\\
&\le 4LC_\epsilon k_n\sqrt{s}\mathbb{E}\left(\sup_{\tau\in\mathcal{T}_n}\max_{j\in J(\delta_0)}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i X_{ij}(1\{Q_i>\tau\}-1\{Q_i>\tau_0\})\right|\right)\\
&\le 4LC_\epsilon k_n\sqrt{s}C_1\left|J(\delta_0)\right|_0\sqrt{\frac{\omega_n^2 s\cdot\log n}{n}},
\end{aligned}
$$

due to the maximal inequality (for VC class indexed by τ and j); when $\delta_0=0$, $V_{1n}\equiv 0$.

$$
\begin{aligned}
V_{2n} &= 4L\mathbb{E}\left(\sup_{|\alpha_J-\alpha_{0J}|_1\le C_\epsilon k_n\sqrt{s}}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i X_{iJ}(\tau_0)^T(\alpha_J-\alpha_{0J})\right|\right)\\
&\le 4LC_\epsilon k_n\sqrt{s}\mathbb{E}\left(\max_{j\in J(\alpha_0)}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i X_{ij}(\tau_0)\right|\right)\le 4LC_\epsilon C_2 k_n^2,
\end{aligned}
$$

due to the Bernstein's moment inequality (Lemma 14.12 of Bühlmann and van de Geer (2011) for some $C_2>0$. Therefore,

$$
V_n\le 4LC_\epsilon k_n\sqrt{s}C_1\left|J(\delta_0)\right|_0\sqrt{\frac{\omega_n^2 s\cdot\log n}{n}}+4LC_\epsilon C_2 k_n^2<5LC_\epsilon C_2 k_n^2,
$$

where the last inequality is due to $C_\epsilon s^3(\log p)^3\log(n)^3=o(n)$. Therefore, conditioning on the event $\widehat{\tau}\in\mathcal{T}_n$ when $\delta_0\ne 0$, or for $\widehat{\tau}\in\mathcal{T}$ when $\delta_0=0$, with probability at least $1-\epsilon$, (2) $\le\frac{1}{\epsilon}5LC_2C_\epsilon k_n^2$.

In addition, note that $P(\max_{j\in J(\alpha_0)}|w_j|=0)=1$, so (3) $=0$ with probability approaching one. Hence

$$\inf_{|\alpha_J-\alpha_{0J}|_2=C_\epsilon k_n} \bar{Q}_n(\alpha_J)-\bar{Q}_n(\alpha_{0J}) \geq \frac{c_1C_\epsilon^2k_n^2}{3}-\frac{1}{\epsilon}5LC_2C_\epsilon k_n^2>0.$$

The last inequality holds for $C_\epsilon > \frac{15LC_2}{c_1\epsilon}$. By the continuity of $\bar{Q}_n$, there is a local minimizer of $\bar{Q}_n(\alpha_J)$ inside $\{\alpha_J \in \mathbb{R}^s : |\alpha_{0J}-\alpha_J|_2 \leq C_\epsilon k_n\}$, which is also a global minimizer due to the convexity. ∎

On $\mathbb{R}^{2p}$, write

$$L_n(\tau,\alpha)=\frac{1}{n}\sum_{i=1}^n \rho(Y_i, X_i(\tau)^T\alpha).$$

For $\bar{\alpha}_J=(\bar{\beta}_{J(\beta_0)},\bar{\delta}_{J(\delta_0)})\equiv(\bar{\beta}_J,\bar{\delta}_J)$ in the previous lemma, define

$$\bar{\alpha}=(\bar{\beta}_J^T,0^T,\bar{\delta}_J^T,0^T)^T.$$

Without introducing confusions, we also write $\bar{\alpha}=(\bar{\alpha}_J,0)$ for notational simplicity. This notation indicates that $\bar{\alpha}$ has zero entries on the indices outside the oracle index set $J(\alpha_0)$. We prove the following lemma.

Lemma B.2. *With probability approaching one, there is a random neighborhood of $\bar{\alpha}$ in $\mathbb{R}^{2p}$, denoted by $\mathcal{H}$, so that $\forall\alpha=(\alpha_J,\alpha_{J^c})\in\mathcal{H}$, if $\alpha_{J^c}\neq 0$, we have $\widetilde{S}_n(\alpha_J,0)<\widetilde{Q}_n(\alpha)$.*

Proof of Lemma B.2. Define an l_2-ball, for $r_n=\mu_n/\log n$,

$$\mathcal{H}=\{\alpha\in\mathbb{R}^{2p}:|\alpha-\bar{\alpha}|_2<r_n/(2p)\}.$$

Then $\sup_{\alpha\in\mathcal{H}}|\alpha-\bar{\alpha}|_1=\sup_{\alpha\in\mathcal{H}}\sum_{l\leq 2p}|\alpha_l-\bar{\alpha}_l|<r_n$. Consider any $\tau\in\mathcal{T}_n$. For any $\alpha=(\alpha_J,\alpha_{J^c})\in\mathcal{H}$, write

$$\begin{aligned}
&L_n(\tau,\alpha_J,0)-L_n(\tau,\alpha)\\
&=L_n(\tau,\alpha_J,0)-\mathbb{E}L_n(\tau,\alpha_J,0)+\mathbb{E}L_n(\tau,\alpha_J,0)-L_n(\tau,\alpha)+\mathbb{E}L_n(\tau,\alpha)-\mathbb{E}L_n(\tau,\alpha)\\
&\leq\mathbb{E}L_n(\tau,\alpha_J,0)-\mathbb{E}L_n(\tau,\alpha)+|L_n(\tau,\alpha_J,0)-\mathbb{E}L_n(\tau,\alpha_J,0)+\mathbb{E}L_n(\tau,\alpha)-L_n(\tau,\alpha)|\\
&\leq\mathbb{E}L_n(\tau,\alpha_J,0)-\mathbb{E}L_n(\tau,\alpha)+|\nu_n(\alpha_J,0,\tau)-\nu_n(\alpha,\tau)|.
\end{aligned}$$

Note that $|(\alpha_J,0)-\bar{\alpha}|_2^2=|\alpha_J-\bar{\alpha}_J|_2^2\leq|\alpha_J-\bar{\alpha}_J|_2^2+|\alpha_{J^c}-0|_2^2=|\alpha-\bar{\alpha}|_2^2$. Hence $\alpha\in\mathcal{H}$ implies $(\alpha_J,0)\in\mathcal{H}$. In addition, by definition of $\bar{\alpha}=(\bar{\alpha}_J,0)$ and $|\bar{\alpha}_J-\alpha_{0J}|_2=O_P(\sqrt{\frac{s\log s}{n}})$

(Lemma B.1), we have $|\bar{\alpha}-\alpha_0|_1 = O_P(s\sqrt{\frac{\log s}{n}})$, which also implies

$$\sup_{\alpha\in\mathcal{H}} |\alpha-\alpha_0|_1 = O_P(s\sqrt{\frac{\log s}{n}}) + r_n,$$

where the randomness in $\sup_{\alpha\in\mathcal{H}} |\alpha-\alpha_0|_1$ comes from that of $\mathcal{H}$.

By the mean value theorem, there is h in the segment between α and $(\alpha_J, 0)$,

$$\begin{aligned}
\mathbb{E}L_n(\tau,\alpha_J,0) - \mathbb{E}L_n(\tau,\alpha) &= \mathbb{E}\rho(Y, X_J(\tau)^T\alpha_J) - \mathbb{E}\rho(Y, X_J(\tau)^T\alpha_J + X_{J^c}(\tau)^T\alpha_{J^c}) \\
&= -\sum_{j\notin J(\alpha_0)} \frac{\partial \mathbb{E}\rho(Y, X(\tau)^T h)}{\partial \alpha_j}\alpha_j \equiv \sum_{j\notin J(\alpha_0)} m_j(\tau,h)\alpha_j
\end{aligned}$$

where $m_j(\tau,h) = -\frac{\partial \mathbb{E}\rho(Y,X(\tau)^T h)}{\partial\alpha_j}$. Hence, $\mathbb{E}L_n(\tau,\alpha_J,0) - \mathbb{E}L_n(\tau,\alpha) \leq \sum_{j\notin J} |m_j(\tau,h)||\alpha_j|$.

Because h is on the segment between α and $(\alpha_J, 0)$, so $h\in\mathcal{H}$. So for all $j\notin J(\alpha_0)$,

$$|m_j(\tau,h)| \leq \sup_{\alpha\in\mathcal{H}} |m_j(\tau,\alpha)| \leq \sup_{\alpha\in\mathcal{H}} |m_j(\tau,\alpha) - m_j(\tau,\alpha_0)| + |m_j(\tau,\alpha_0) - m_j(\tau_0,\alpha_0)|.$$

We now argue that we can apply Assumption 4.1 (ii). Let

$$c_n = s\sqrt{(\log s)/n} + r_n.$$

For any $\epsilon>0$, there is $C_\epsilon>0$, with probability at last $1-\epsilon$, $\sup_{\alpha\in\mathcal{H}} |\alpha-\alpha_0|_1 \leq C_\epsilon c_n$. $\forall \alpha\in\mathcal{H}$, write $\alpha = (\beta,\delta)$ and $\theta = \beta+\delta$. On the event $|\alpha-\alpha_0|_1 \leq C_\epsilon c_n$, we have $|\beta-\beta_0|_1 \leq C_\epsilon c_n$ and $|\theta-\theta_0|_1 \leq C_\epsilon c_n$. Hence $\mathbb{E}[(X^T(\beta-\beta_0))^2 1\{Q\leq\tau_0\}] \leq |\beta-\beta_0|_1^2 \max_{i,j\leq p} E|X_iX_j| < r^2$, yielding $\beta\in\mathcal{B}(\beta_0,r)$. Similarly, $\theta\in\mathcal{G}(\theta_0,r)$. Therefore, by Assumption 4.1 (ii), with probability at least $1-\epsilon$, (note that neither C_ϵ, L nor c_n depend on α)

$$\begin{aligned}
\max_{j\notin J(\alpha_0)} \sup_{\tau\in\mathcal{T}_n} \sup_{\alpha\in\mathcal{H}} |m_j(\tau,\alpha) - m_j(\tau,\alpha_0)| &\leq L \sup_{\alpha\in\mathcal{H}} |\alpha-\alpha_0|_1 \leq L(C_\epsilon c_n), \\
\max_{j\leq 2p} \sup_{\tau\in\mathcal{T}_n} |m_j(\tau,\alpha_0) - m_j(\tau_0,\alpha_0)| &\leq M_n\omega_n^2 s\cdot\log n.
\end{aligned}$$

In particular, when $\delta_0 = 0$, $m_j(\tau,\alpha_0) = 0$ for all τ. Therefore, when $\delta_0\neq 0$,

$$\sup_{j\notin J(\alpha_0)} \sup_{\tau\in\mathcal{T}_n} |m_j(\tau,h)| = O_P(c_n + M_n\omega_n^2 s\cdot\log n) = o_P(\mu_n);$$

when $\delta_0 = 0$, $\sup_{j\notin J(\alpha_0)} \sup_{\tau\in\mathcal{T}} |m_j(\tau,h)| = O_P(c_n) = o_P(\mu_n)$.

Let $\epsilon_1, ..., \epsilon_n$ be a Rademacher sequence independent of $\{Y_i, X_i, Q_i\}_{i\leq n}$. Then by the symmetrization and contraction theorems,

$$
\begin{aligned}
&\mathbb{E}\left(\sup_{\tau\in\mathcal{T}}|\nu_n(\alpha_J,0,\tau)-\nu_n(\alpha,\tau)|\right)\\
&\leq 2\mathbb{E}\left(\sup_{\tau\in\mathcal{T}}\left|\frac{1}{n}\sum_{i=1}^n\epsilon_i[\rho(Y_i,X_{iJ}(\tau)^T\alpha_J)-\rho(Y_i,X_i(\tau)^T\alpha)]\right|\right)\\
&\leq 4L\mathbb{E}\left(\sup_{\tau\in\mathcal{T}}\left|\frac{1}{n}\sum_{i=1}^n\epsilon_i[X_{iJ}(\tau)^T\alpha_J-X_i(\tau)^T\alpha]\right|\right)\\
&\leq 4L\mathbb{E}\left(\sup_{\tau\in\mathcal{T}}\left\|\frac{1}{n}\sum_{i=1}^n\epsilon_iX_i(\tau)\right\|_{\max}\right)\sum_{j\notin J(\alpha_0)}|\alpha_j|\leq 2\omega_n\sum_{j\notin J(\alpha_0)}|\alpha_j|,
\end{aligned}
$$

where the last equality follows from (A.5).

Thus uniformly over $\alpha\in\mathcal{H}$, $L_n(\tau,\alpha_J,0)-L_n(\tau,\alpha)=o_P(\mu_n)\sum_{j\notin J(\alpha_0)}|\alpha_j|$. On the other hand,

$$
\sum_{j\in J(\alpha_0)}w_j\mu_n\widehat{D}_j|\alpha_j|-\sum_j w_j\mu_n\widehat{D}_j|\alpha_j|=\sum_{j\notin J(\alpha_0)}\mu_n w_j\widehat{D}_j|\alpha_j|.
$$

Also, with probability approaching one, $w_j=1$ and $\widehat{D}_j\geq\overline{D}$ for all $j\notin J(\alpha_0)$. Hence with probability approaching one, $\widetilde{Q}_n(\alpha_J,0)-\widetilde{Q}_n(\alpha)$ equals

$$
L_n(\widehat{\tau},\alpha_J,0)+\sum_{j\in J(\alpha_0)}\widehat{D}_jw_j\lambda_n|\alpha_j|-L_n(\widehat{\tau},\alpha)-\sum_{j\leq 2p}\widehat{D}_jw_j\omega_n|\alpha_j|\leq-\underline{D}\frac{\mu_n}{2}\sum_{j\notin J(\alpha_0)}|\alpha_j|<0. \blacksquare
$$

Proof of Theorems 4.1 and 4.2. By Lemmas B.1 and B.2, with probability approaching one, for any $\alpha=(\alpha_J,\alpha_{J^c})\in\mathcal{H}$,

$$
\widetilde{S}_n(\bar{\alpha}_J,0)=\bar{Q}_n(\bar{\alpha}_J)\leq\bar{Q}_n(\alpha_J)=\widetilde{S}_n(\alpha_J,0)\leq\widetilde{S}_n(\alpha).
$$

Hence $(\bar{\alpha}_J,0)$ is a local minimizer of $\widetilde{S}_n$, which is also a global minimizer due to the convexity. This implies that with probability approaching one, $\widetilde{\alpha}=(\widetilde{\alpha}_J,\widetilde{\alpha}_{J^c})$ satisfies: $\widetilde{\alpha}_{J^c}=0$, and $\widetilde{\alpha}_J=\bar{\alpha}_J$, so

$$
|\widetilde{\alpha}_J-\alpha_{0J}|_2=O_P\left(\sqrt{\frac{s\log s}{n}}\right),\quad|\widetilde{\alpha}_J-\alpha_{0J}|_1=O_P\left(s\sqrt{\frac{\log s}{n}}\right).
$$

$\blacksquare$

Proof of Lemma 4.1. Noting that

$$
\rho\left(Y_i,X_i^T\beta+X_i^T\delta 1\{Q_i>\tau\}\right)=\rho\left(Y_i,X_i^T\beta\right)1\{Q_i\leq\tau\}+\rho\left(Y_i,X_i^T\beta+X_i^T\delta\right)1\{Q_i>\tau\},
$$

we have, for $\tau > \tau_0$,

$$\begin{aligned}
&D_n(\alpha,\tau) \\
&\equiv \{Q_n^*(\alpha,\tau) - Q_n^*(\alpha,\tau_0)\} - \{Q_n^*(\alpha_0,\tau) - Q_n^*(\alpha_0,\tau_0)\} \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[\rho\left(Y_i, X_i^T\beta\right) - \rho\left(Y_i, X_i^T\beta_0\right)\right] 1\{\tau_0 < Q_i \le \tau\} \\
&- \frac{1}{n}\sum_{i=1}^{n}\left[\rho\left(Y_i, X_i^T\beta + X_i^T\delta\right) - \rho\left(Y_i, X_i^T\beta_0 + X_i^T\delta_0\right)\right] 1\{\tau_0 < Q_i \le \tau\} \\
&=: D_{n1}(\alpha,\tau) - D_{n2}(\alpha,\tau).
\end{aligned}$$

To prove this lemma, we consider empirical processes

$$\mathbb{G}_{nj}(\alpha_J,\tau) \equiv \sqrt{n}\left(D_{nj}(\alpha_J,\tau) - \mathbb{E}D_{nj}(\alpha_J,\tau)\right), \quad (j = 1, 2),$$

and apply the maximal inequality in Theorem 2.14.2 of van der Vaart and Wellner (1996).

First, for $\mathbb{G}_{n1}(\alpha_J,\tau)$, we consider the following class of functions indexed by (β_J,τ):

$$\mathcal{F}_n \equiv \left\{\left(\rho\left(Y_i, X_{iJ}^T\beta_J\right) - \rho\left(Y_i, X_{iJ}^T\beta_{0J}\right)\right) 1(\tau_0 < Q_i \le \tau) : |\beta_J - \beta_{0J}|_2 \le Kr_n \text{ and } |\tau - \tau_0| \le Ks_n\right\}.$$

Note that the Lipschitz property of ρ yields that

$$\left|\rho\left(Y_i, X_{iJ}^T\beta_J\right) - \rho\left(Y_i, X_{iJ}^T\beta_{0J}\right)\right| 1\{\tau_0 < Q_i \le \tau\} \le \left|X_{iJ}^T\right|_2 |\beta_J - \beta_{0J}|_2 1\{|Q_i - \tau_0| \le Ks_n\}.$$

Thus, we let the envelope function be $F_n(X_{iJ}, Q_i) \equiv |X_{iJ}|_2 Kr_n 1\{|Q_i - \tau_0| \le Ks_n\}$ and note that its L_2 norm is $O\left(\sqrt{s}r_n\sqrt{s_n}\right)$.

To compute the bracketing integral

$$J_{[]}(1, \mathcal{F}_n, L_2) \equiv \int_0^1 \sqrt{1 + \log N_{[]}(\varepsilon\|F_n\|_{L_2}, \mathcal{F}_n, L_2)}\,d\varepsilon,$$

note that its 2ε bracketing number is bounded by the product of the ε bracketing numbers of two classes $\mathcal{F}_{n1} \equiv \left\{\rho\left(Y_i, X_{iJ}^T\beta_J\right) - \rho\left(Y_i, X_{iJ}^T\beta_0\right) : |\beta_J - \beta_{0J}|_2 \le Kr_n\right\}$ and $\mathcal{F}_{n2} \equiv \{1(\tau_0 < Q_i \le \tau) : |\tau - \tau_0| \le Ks_n\}$ by Lemma 9.25 of Kosorok (2008) since both classes are bounded w.p.a.1 (note that w.p.a.1, $|X_{iJ}|_2 Kr_n < C < \infty$ for some constant C). That is,

$$N_{[]}(2\varepsilon\|F_n\|_{L_2}, \mathcal{F}_n, L_2) \le N_{[]}(\varepsilon\|F_n\|_{L_2}, \mathcal{F}_{n1}, L_2)\, N_{[]}(\varepsilon\|F_n\|_{L_2}, \mathcal{F}_{n2}, L_2).$$

Let $F_{n1}(X_{iJ}) \equiv |X_{iJ}|_2 Kr_n$ and $l_n(X_{iJ}) \equiv |X_{iJ}|_2$. Note that by Theorem 2.7.11 of van der

Vaart and Wellner (1996), the Lipschitz property of ρ implies that

$$N_{[]}\left(2\varepsilon\|l_n\|_{L_2}, \mathcal{F}_{n1}, L_2\right) \leq N(\varepsilon, \{\beta_J : |\beta_J - \beta_{0J}|_2 \leq Kr_n\}, |\cdot|_2),$$

which in turn implies that, for some constant C,

$$\begin{aligned} N_{[]}\left(\varepsilon\|F_n\|_{L_2}, \mathcal{F}_{n1}, L_2\right) &\leq N\left(\frac{\varepsilon\|F_n\|_{L_2}}{2\|l_n\|_{L_2}}, \{\beta_J : |\beta_J - \beta_{0J}|_2 \leq Kr_n\}, |\cdot|_2\right) \\ &\leq C\left(\frac{\sqrt{s}}{\varepsilon\sqrt{s_n}}\right)^s = C\left(\frac{\sqrt{n}}{\varepsilon(\log p)^{3/2}(\log n)}\right)^s, \end{aligned}$$

where the last inequality holds since a ε-ball contains a hypercube with side length $\varepsilon/\sqrt{s}$ in the s-dimensional Euclidean space. On the other hand, for the second class of functions $\mathcal{F}_{n2}$ with the envelope function $F_{n2}(Q_i) \equiv 1\{|Q_i - \tau_0| \leq Ks_n\}$, we have that

$$N_{[]}\left(\varepsilon\|F_n\|_{L_2}, \mathcal{F}_{n2}, L_2\right) \leq C\frac{\sqrt{s_n}}{\varepsilon\|F_n\|_{L_2}} = \frac{C}{\varepsilon\sqrt{s}r_n} = \frac{C\sqrt{n}}{\varepsilon s\sqrt{\log s}},$$

for some constant C. Combining these results together yields that

$$N_{[]}\left(\varepsilon\|F_n\|_{L_2}, \mathcal{F}_n, L_2\right) \leq \frac{C^2\sqrt{n}}{\varepsilon s\sqrt{\log s}}\left(\frac{\sqrt{n}}{\varepsilon(\log p)^{3/2}(\log n)}\right)^s \leq C^2\varepsilon^{-s-1}n^{s/2}$$

for all sufficiently large n. Then we have that

$$J_{[]}\left(1, \mathcal{F}_n, L_2\right) \leq C^2(\sqrt{s\log n} + \sqrt{s})$$

for all sufficiently large n. Thus, by the maximal inequality in Theorem 2.14.2 of van der Vaart and Wellner (1996),

$$\begin{aligned} n^{-1/2}\,\mathbb{E}\sup_{\mathcal{A}_n\times\mathcal{T}_n}\left|\mathbb{G}_{n1}\left(\alpha_J, \tau\right)\right| &\leq O\left[n^{-1/2}\sqrt{s}r_n\sqrt{s_n}(\sqrt{s\log n} + \sqrt{s})\right] \\ &= O\left[\frac{s^{3/2}}{n^{3/2}}\sqrt{\log s}(\log p)^{3/2}(\log n)(\sqrt{s\log n} + \sqrt{s})\right] \\ &= o\left(n^{-1}\right), \end{aligned}$$

where the last equality follows from the restriction that $s^4(\log s)(\log p)^3(\log n)^3 = o(n)$. Identical arguments also apply to $\mathbb{G}_{n2}\left(\alpha_J, \tau\right)$.

Turning to $\mathbb{E}D_n\left(\alpha, \tau\right)$, note that by the condition that $\frac{\partial}{\partial\alpha}E\left[\rho\left(Y, X^T\alpha\right)|Q = t\right]$ exists for all t in a neighborhood of τ_0 and all its elements are continuous and bounded below and above, we have that for some mean value $\tilde{\beta}_J$ between β_J and β_{0J},

$$\begin{aligned}
&\left|\mathbb{E}\left(\rho\left(Y, X_J^T\beta_J\right) - \rho\left(Y, X_J^T\beta_{0J}\right)\right) 1\left\{\tau_0 < Q \leq \tau\right\}\right| \\
&= \left|\mathbb{E}\left[\frac{\partial}{\partial\beta}\mathbb{E}\left[\rho\left(Y, X^T\tilde{\beta}_J\right)|Q\right] 1\left\{\tau_0 < Q \leq \tau\right\}\right](\beta - \beta_0)\right| \\
&= O\left(sr_n s_n\right) \\
&= O\left[\frac{s^{5/2}}{n^{3/2}}\sqrt{\log s}(\log p)^3(\log n)^2\right] \\
&= o\left(n^{-1}\right),
\end{aligned}$$

where the last equality follows from the restriction that $s^5(\log s)(\log p)^6(\log n)^4 = o(n)$. Since the same holds for the other term in $\mathbb{E}D_n$, $\sup|\mathbb{E}D_n(\alpha,\tau)| = o(n^{-1})$ as desired. ∎

C Proofs of Section 5

C.1 Proof of Lemma 5.1

Verification of Assumption 3.3 (i). The loss function for quantile regression is convex and satisfies the Liptschitz condition. ∎

Verification of Assumption 3.3 (ii). Note that $\rho(Y,t) = h_\gamma(Y-t)$, where $h_\gamma(t) = t(\gamma - 1\{t \leq 0\})$. By (B.3) of Belloni and Chernozhukov (2011),

$$h_\gamma(w-v) - h_\gamma(w) = -v(\gamma - 1\{w \leq 0\}) + \int_0^v (1\{w \leq z\} - 1\{w \leq 0\})dz \tag{C.1}$$

where $w = Y - X(\tau_0)^T\alpha_0$ and $v = X(\tau_0)^T(\alpha - \alpha_0)$. Note that

$$\mathbb{E}[v(\gamma - 1\{w \leq 0\})|Q] = -\mathbb{E}[X(\tau_0)^T(\alpha - \alpha_0)(\gamma - 1\{U \leq 0\})|Q] = 0,$$

since $\mathbb{P}(U \leq 0|X,Q) = \gamma$. Let $F_{Y|X,Q}$ denote the CDF of the conditional distribution $Y|X,Q$.

Then

$$
\begin{aligned}
&\mathbb{E}\left[\rho(Y, X(\tau_0)^T\alpha) - \rho(Y, X(\tau_0)^T\alpha_0)|Q\right] \\
&= \mathbb{E}\left[\int_0^{X(\tau_0)^T(\alpha-\alpha_0)} (1\{U \le z) - 1\{U \le 0\})dz \middle| Q\right] \\
&= \mathbb{E}\left[\int_0^{X(\tau_0)^T(\alpha-\alpha_0)} [F_{Y|X,Q}(X(\tau_0)^T\alpha_0 + z|X,Q) - F_{Y|X,Q}(X(\tau_0)^T\alpha_0|X,Q)]dz \middle| Q\right] \\
&\ge 0,
\end{aligned}
$$

where the last inequality follows immediately from the fact that $F_{Y|X,Q}(\cdot|X,Q)$ is the CDF. Hence, we have verified Assumption 3.3 (ii). ∎

Verification of Assumption 3.3 (iii). Following the arguments analogous those used in (B.4) of Belloni and Chernozhukov (2011), the mean value expansion implies:

$$
\begin{aligned}
&\mathbb{E}\left[\rho(Y, X(\tau_0)^T\alpha) - \rho(Y, X(\tau_0)^T\alpha_0)|Q\right] \\
&= \mathbb{E}\left\{\int_0^{X(\tau_0)^T(\alpha-\alpha_0)} \left[z f_{Y|X,Q}(X(\tau_0)^T\alpha_0|X,Q) + \frac{z^2}{2}\tilde{f}_{Y|X,Q}(X(\tau_0)^T\alpha_0 + t|X,Q)\right] dz \middle| Q\right\} \\
&= \frac{1}{2}(\alpha-\alpha_0)^T\mathbb{E}\left[X(\tau_0)X(\tau_0)^T f_{Y|X,Q}(X(\tau_0)^T\alpha_0|X,Q)|Q\right](\alpha-\alpha_0) \\
&+\mathbb{E}\left\{\int_0^{X(\tau_0)^T(\alpha-\alpha_0)} \frac{z^2}{2}\tilde{f}_{Y|X,Q}(X(\tau_0)^T\alpha_0 + t|X,Q)dz \middle| Q\right\}
\end{aligned}
$$

for some intermediate value t between 0 and z. By condition (ii) of Assumption 5.1,

$$
|\tilde{f}_{Y|X,Q}(X(\tau_0)^T\alpha_0 + t|X,Q)| \le C_1 \quad \text{and} \quad f_{Y|X,Q}(X(\tau_0)^T\alpha_0|X,Q) \ge C_2.
$$

Hence, taking the expectation on $\{Q \le \tau_0\}$ gives

$$
\begin{aligned}
&\mathbb{E}\left[\rho(Y, X^T\beta) - \rho(Y, X^T\beta_0)1\{Q \le \tau_0\}\right] \\
&\ge \frac{C_2}{2}\mathbb{E}[(X^T(\beta-\beta_0))^2 1\{Q \le \tau_0\}] - \frac{C_1}{6}\mathbb{E}[(X^T(\beta-\beta_0))^3 1\{Q \le \tau_0\}] \\
&\ge \frac{C_2}{4}\mathbb{E}[|X^T(\beta-\beta_0)|^2 1\{Q \le \tau_0\}],
\end{aligned}
$$

where the last inequality follows from

$$
\frac{C_2}{4}\mathbb{E}[|X^T(\beta-\beta_0)|^2 1\{Q \le \tau_0\}] \ge \frac{C_1}{6}\mathbb{E}[|X^T(\beta-\beta_0)|^3 1\{Q \le \tau_0\}]. \tag{C.2}
$$

To see why (C.2) holds, note that by (5.1), for any nonzero $\beta \in \mathcal{B}(\beta_0, r^*_{QR})$,

$$\frac{\mathbb{E}[|X^T(\beta-\beta_0)|^2 1\{Q \le \tau_0\}]^{3/2}}{\mathbb{E}[|X^T(\beta-\beta_0)|^3 1\{Q \le \tau_0\}]} \ge r^*_{QR}\frac{2C_1}{3C_2} \ge \frac{2C_1}{3C_2}\mathbb{E}[|X^T(\beta-\beta_0)|^2 1\{Q \le \tau_0\}]^{1/2},$$

which proves (C.2) immediately. Thus, we have shown that Assumption 3.3 (iii) holds for $r_1(\eta)$ with $\eta^* = C_2/4$ and $r^* = r^*_{QR}$ defined in (5.1) in Assumption 5.1. The case for $r_2(\eta)$ is similar and hence is omitted. ∎

Verification of Assumption 3.3 (iv). We again start from (C.1) but with different choices of (w, v) such that $w = Y - X(\tau_0)^T\alpha_0$ and $v = X^T\delta_0[1\{Q \le \tau_0\} - 1\{Q > \tau_0\}]$. Then arguments similar to those used in verifying Assumptions 3.3 (ii)-(iii) yield that for $\tau < \tau_0$,

$$\mathbb{E}\left[\rho\left(Y, X^T\theta_0\right) - \rho\left(Y, X^T\beta_0\right)|Q=\tau\right] \tag{C.3}$$

$$= \mathbb{E}\left\{\int_0^{X^T\delta_0} z f_{Y|X,Q}(X^T\beta_0 + t|X,Q)dz\middle|Q=\tau\right\} \tag{C.4}$$

$$\ge \mathbb{E}\left\{\int_0^{\widetilde{\varepsilon}(X^T\delta_0)} z f_{Y|X,Q}(X^T\beta_0 + t|X,Q)dz\middle|Q=\tau\right\} \tag{C.5}$$

$$\ge \frac{\widetilde{\varepsilon}^2 C_3}{2}\mathbb{E}\left[(X^T\delta_0)^2|Q=\tau\right], \tag{C.6}$$

where t is an intermediate value t between 0 and z. Here, if the extra condition such that $M_3^{-1} < \mathbb{E}[(X^T\delta_0)^2|Q=\tau] \le M_3$ for some $M_3 > 0$ does not hold, we need to rely on (5.3) in Assumption 5.2 to prove the last inequality in (C.3). Thus, we have that

$$\mathbb{E}\left[\left(\rho\left(Y, X^T\theta_0\right) - \rho\left(Y, X^T\beta_0\right)\right)1\{\tau < Q \le \tau_0\}\right] \ge \frac{\widetilde{\varepsilon}^2 C_3}{2}\mathbb{E}\left[\left(X^T(\beta_0 - \theta_0)\right)^2 1\{\tau < Q \le \tau_0\}\right].$$

The case that $\tau > \tau_0$ is similar. ∎

Verification of Assumption 4.1. Recall that $m_j(\tau,\alpha) = \mathbb{E}[X_j(\tau)(1\{Y - X(\tau)^T\alpha \le 0\} - \gamma)]$. Hence, note that $m_j(\tau_0, \alpha_0) = 0$, for all $j \le 2p$. For condition (i) of Assumption 4.1, for all

$j \leq 2p$,

$$
\begin{aligned}
&|m_j(\tau,\alpha_0) - m_j(\tau_0,\alpha_0)| \\
&= |\mathbb{E}X_j(\tau)[1\{Y \leq X(\tau)^T\alpha_0\} - 1\{Y \leq X(\tau_0)^T\alpha_0\}]| \\
&= |\mathbb{E}X_j(\tau)[\mathbb{P}(Y \leq X(\tau)^T\alpha_0|X,Q) - \mathbb{P}(Y \leq X(\tau_0)^T\alpha_0|X,Q)]| \\
&\leq C\mathbb{E}|X_j(\tau)||(X(\tau) - X(\tau_0))^T\alpha_0| \\
&= C\mathbb{E}|X_j(\tau)||X^T\delta_0(1\{Q > \tau\} - 1\{Q > \tau_0\})| \\
&\leq C\mathbb{E}|X_j(\tau)||X^T\delta_0|(1\{\tau < Q < \tau_0\} + 1\{\tau_0 < Q < \tau\}) \\
&\leq C(\mathbb{P}(\tau_0 < Q < \tau) + \mathbb{P}(\tau < Q < \tau_0)) \sup_{\tau,\tau' \in \mathcal{T}_0} \mathbb{E}(|X_j(\tau)X^T\delta_0||Q = \tau') \\
&\leq C(\mathbb{P}(\tau_0 < Q < \tau) + \mathbb{P}(\tau < Q < \tau_0)) \sup_{\tau,\tau' \in \mathcal{T}_0} [\mathbb{E}(|X_j(\tau)|^2||Q = \tau')]^{1/2}[\mathbb{E}(|X^T\delta_0|^2|Q = \tau')]^{1/2} \\
&\leq CM_2K_2|\delta_0|_2|\tau_0 - \tau|
\end{aligned}
$$

for some constant C, where the last inequality follows from Assumptions 3.1 (i), (iv), 3.2 (i), and 3.4 (i). Therefore, we have verified condition (i) of Assumption 4.1 with $M_n = CM_2K_2|\delta_0|_2$.

We now verify condition (ii) of Assumption 4.1. For all j and τ in a neighborhood of τ_0,

$$
\begin{aligned}
&|m_j(\tau,\alpha) - m_j(\tau,\alpha_0)| = |\mathbb{E}X_j(\tau)(1\{Y \leq X(\tau)^T\alpha\} - 1\{Y \leq X(\tau)^T\alpha_0\})| \\
&= |\mathbb{E}X_j(\tau)(\mathbb{P}(Y \leq X(\tau)^T\alpha|X,Q) - \mathbb{P}(Y \leq X(\tau)^T\alpha_0|X,Q))| \\
&\leq C\mathbb{E}|X_j(\tau)||X(\tau)^T(\alpha - \alpha_0)| \leq C|\alpha - \alpha_0|_1 \max_{j \leq 2p, i \leq 2p} \mathbb{E}|X_j(\tau)X_i(\tau)|,
\end{aligned}
$$

which implies the result immediately in view of Assumption 3.1 (iv). Finally, it is straightforward to verify condition (iii). ∎

C.2 Proof of Lemma 5.2

We shall let $C > 0$ denote a generic constant.

Verification of Assumption 3.3 (i). The loss function for logistic regression is convex and satisfies the Liptschitz condition. ∎

Verification of Assumption 3.3 (ii). Recall that $g(t) = \exp(t)/(1 + \exp(t))$; then for all α,

$$
\mathbb{E}[\rho(Y, X(\tau_0)^T\alpha)|Q] = \mathbb{E}[f(g(X(\tau_0)^T\alpha), t_0)|Q], \quad f(t, t_0) = -t_0 \log t - (1 - t_0)\log(1 - t),
$$

where $t_0 = g(X(\tau_0)^T\alpha_0)$. Note that $f(t, t_0) \geq f(t_0, t_0)$ for all $t > 0$. Hence, we have verified the assumption. ∎

Verification of Assumption 3.3 (iii). Note that $\forall\beta \in \mathcal{B}(\beta_0, r)$,

$$\mathbb{E}\left(\rho\left(Y, X^T\beta\right) 1\left\{Q \le \tau_0\right\}\right) = \mathbb{E}\left(f(g(X^T\beta), g(X^T\beta_0)) 1\left\{Q \le \tau_0\right\}\right).$$

Let $t_0 = g(X^T\beta_0)$, then $\partial_t f(t, t_0)|_{t=t_0} = 0$. Let $t = g(X^T\beta)$. By Taylor's expansion, there are $\lambda \in [0, 1]$ and $\tilde{t} \in (0, 1)$ such that $f(t, t_0) - f(t_0, t_0) = \partial_t^2 f(\tilde{t}, t_0)(t - t_0)^2/2$, which implies, for $\tilde{\beta} = \lambda\beta + (1 - \lambda)\beta_0$,

$$\mathbb{E}[\{\rho(Y, X^T\beta) - \rho(Y, X^T\beta_0)\}1\{Q \le \tau_0\}] = \frac{1}{2}\mathbb{E}[\partial_t^2 f(\tilde{t}, t_0) g'(X^T\tilde{\beta})^2(X^T\beta - X^T\beta_0)^2 1\{Q \le \tau_0\}].$$

By Assumption 5.3, $\partial_t^2 f(\tilde{t}, t_0) = t_0/\tilde{t}^2 + (1 - t_0)/(1 - \tilde{t})^2 > C$ and $\epsilon < g(X^T\tilde{\beta}) < 1 - \epsilon$, so $g'(X^T\tilde{\beta})^2 = g(X^T\tilde{\beta})(1 - g(X^T\tilde{\beta})) > \epsilon^2$. Hence

$$\mathbb{E}[\{\rho(Y, X^T\beta) - \rho(Y, X^T\beta_0)\}1\{Q \le \tau_0\} \ge \frac{C\epsilon^2}{2}\mathbb{E}[(X^T\beta - X^T\beta_0)^2 1\{Q \le \tau_0\}].$$

So the assumption holds with $\eta^* = C\epsilon^2/2$. The inequality $r_2(\eta^*) \ge r^*$ can be proved using the same argument. ∎

Verification of Assumption 3.3 (iv). Note that for all $\tau > \tau_0$, note that for $t_0 = g(X^T\theta_0)$, and $t = g(X^T\beta_0)$,

$$\mathbb{E}\left[\rho(Y, X^T\beta_0) - \rho(Y, X^T\theta_0)|Q = \tau\right] = \mathbb{E}[f(t, t_0) - f(t_0, t_0)|Q = \tau].$$

Using the same argument as verifying Assumption 3.3 (ii)-(iii), there exists a $C > 0$ such that the right hand side is bounded below by $C\mathbb{E}[g'(X^T\tilde{\beta})^2(X^T\delta_0)^2|Q = \tau]$, where for some $\lambda > 0$, $\tilde{\beta} = \lambda\beta_0 + (1 - \lambda)\theta_0$. We now consider lower bound $g'(X^T\tilde{\beta})$. By Assumption 5.3, almost surely there is $\epsilon > 0$, $\epsilon < g(X^T\beta_0), g(X^T\theta_0) < 1 - \epsilon$. By the monotonicity of $g(t)$, and $\min\{X\beta_0, X\theta_0\} \le X^T\tilde{\beta} \le \max\{X\beta_0, X\theta_0\}$, we have $\epsilon < g(X^T\tilde{\beta}) < 1 - \epsilon$. Moreover, $g'(t) = g(t)(1 - g(t))$, so $g'(X^T\tilde{\beta})^2 > \epsilon^4$.

In addition, $\inf_{\tau > \tau_0} \mathbb{E}[(X^T\delta_0)^2|Q = \tau]$ is bounded away from zero. Hence there is $C^* > 0$, $\mathbb{E}\left[\rho(Y, X^T\beta_0) - \rho(Y, X^T\theta_0)|Q = \tau\right] > C^*$. Therefore, we have, for some $C > 0$,

$$\mathbb{E}\left[\left(\rho\left(Y, X^T\theta_0\right) - \rho\left(Y, X^T\beta_0\right)\right) 1\left\{\tau < Q \le \tau_0\right\}\right] \ge C\mathbb{E}\left[(X^T\delta_0)^2 1\left\{\tau < Q \le \tau_0\right\}\right].$$

The case of $\tau < \tau_0$ is similar. ∎

Verification of Assumption 4.1. For part (i) of Assumption 4.1, by the mean value theorem, for all $j \leq 2p$,

$$
\begin{aligned}
&|m_j(\tau,\alpha_0) - m_j(\tau_0,\alpha_0)| = \left|\mathbb{E}\left\{\frac{g(X(\tau_0)^T\alpha_0) - g(X(\tau)^T\alpha_0)}{g(X(\tau)^T\alpha_0)(1-g(X(\tau)^T\alpha_0))} g'(X(\tau)^T\alpha)X_j(\tau)\right\}\right| \\
&\leq C\sup_t |g'(t)|^2 \mathbb{E}|X^T\delta_0(1\{Q>\tau_0\} - 1\{Q>\tau\})X_j(\tau)| \\
&\leq (\mathbb{P}(\tau_0 < Q < \tau) + \mathbb{P}(\tau < Q < \tau_0)) \sup_{\tau,\tau'\in\mathcal{T}_0} \mathbb{E}(|X_j(\tau)X^T\delta_0||Q=\tau') \\
&\leq C(\mathbb{P}(\tau_0 < Q < \tau) + \mathbb{P}(\tau < Q < \tau_0)) \sup_{\tau,\tau'\in\mathcal{T}_0} [\mathbb{E}(|X_j(\tau)|^2||Q=\tau')]^{1/2}[\mathbb{E}(|X^T\delta_0|^2|Q=\tau')]^{1/2} \\
&\leq CM_2K_2|\delta_0|_2|\tau_0-\tau|
\end{aligned}
$$

for some constant C, where the last inequality follows from Assumptions 3.1 (i), (iv), 3.2 (i), and 3.4 (i). Therefore, we have verified condition (i) of Assumption 4.1 with $M_n = CM_2K_2|\delta_0|_2$.

For part (ii), since $\epsilon < g(X(\tau)^T\alpha) < 1-\epsilon$, by the mean value theorem, there is Z,

$$
\begin{aligned}
&|m_j(\tau,\alpha) - m_j(\tau,\alpha_0)| \leq |\mathbb{E}\left\{g(X(\tau_0)^T\alpha_0)\frac{g(X(\tau)^T\alpha_0) - g(X(\tau)^T\alpha)}{g(X(\tau)^T\alpha)g(X(\tau)^T\alpha_0)} g'(X(\tau)^T\alpha)X_j(\tau)\right\}| \\
&+|\mathbb{E}\left\{(1-g(X(\tau_0)^T\alpha_0))\frac{g(X(\tau)^T\alpha) - g(X(\tau)^T\alpha_0)}{(1-g(X(\tau)^T\alpha))(1-g(X(\tau)^T\alpha_0))} g'(X(\tau)^T\alpha)X_j(\tau)\right\}| \\
&+|\mathbb{E}\left\{\left[\frac{g(X(\tau_0)^T\alpha_0)}{g(X(\tau)^T\alpha_0)} - \frac{1-g(X(\tau_0)^T\alpha_0)}{1-g(X(\tau)^T\alpha_0)}\right](g'(X(\tau)^T\alpha_0) - g'(X(\tau)^T\alpha))X_j(\tau)\right\}| \\
&\leq C\max_{j,m\leq 2p} \mathbb{E}|X_j(\tau)X_m(\tau)||\alpha-\alpha_0|_1 \\
&+|\mathbb{E}\left\{\left[\frac{g(X(\tau_0)^T\alpha_0)}{g(X(\tau)^T\alpha_0)} - \frac{1-g(X(\tau_0)^T\alpha_0)}{1-g(X(\tau)^T\alpha_0)}\right] g''(Z)X_j(\tau)X(\tau)^T(\alpha_0-\alpha)\right\}| \\
&\leq C\max_{j,m\leq 2p} \mathbb{E}|X_j(\tau)X_m(\tau)||\alpha_0-\alpha|_1,
\end{aligned}
$$

which implies the result immediately in view of Assumption 3.1 (iv). Finally, condition (iii) can be verified by straightforward calculations. ∎

References

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

Belloni, A., Chernozhukov, V. and Wang, L. (2011). Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika* **98** 791–806.

Bickel, P., Ritov, Y. and Tsybakov, A. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics* **37** 1705–1732.

Bradic, J., Fan, J. and Wang, W. (2011). Penalized composite quasi-likelihood for ultrahigh dimensional variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **73** 325–349.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for high-dimensional data, methods, theory and applications*. Springer, New York.

Candes, E. and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Annals of Statistics* **35** 2313–2351.

Chan, K.-S. (1993). Consistency and limiting distribution of the least squares estimator of a threshold autoregressive model. *Annals of Statistics* **21** 520–533.

Chan, N. H., Yau, C. Y. and Zhang, R.-M. (2014). Group LASSO for structural break time series. *Journal of the American Statistical Association* **109** 590–599.

Cho, H. and Fryzlewicz, P. (2014). Multiple-change-point detection for high dimensional time series via sparsified binary segmentation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* forthcoming.

Ciuperca, G. (2013). Quantile regression in high-dimension with breaking. *arXiv preprint* http://arxiv.org/abs/1302.4244.

Enikeeva, F. and Harchaoui, Z. (2013). High-dimensional change-point detection with sparse alternatives. *arXiv preprint* http://arxiv.org/abs/1312.1900.

Fan, J., Fan, Y. and Barut, E. (2014a). Adaptive robust variable selection. *Annals of Statistics* **42** 324–351.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96** 1348–1360.

Fan, J., Xue, L. and Zou, H. (2014b). Strong oracle optimality of folded concave penalized estimation. *Annals of Statistics* **42** 819–849.

Frick, K., Munk, A. and Sieling, H. (2014). Multiscale change point inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76** 495–580.

Hansen, B. E. (2000). Sample splitting and threshold estimation. *Econometrica* **68** 575–603.

He, X. and Shao, Q.-M. (2000). On parameters of increasing dimensions. *Journal of Multivariate Analysis* **73** 120–135.

Johnstone, I. (1994). On minimax estimation of a sparse normal mean vector. *Annals of Statistics* **22** 271–289.

Kosorok, M. R. (2008). *Introduction to Empirical Processes and Semiparametric Inference*. Springer, New York.

KOSOROK, M. R. and SONG, R. (2007). Inference under right censoring for transformation models with a change-point based on a covariate threshold. *Annals of Statistics* **35** 957–989.

LAWLEY, D. and MAXWELL, A. (1971). *Factor analysis as a statistical method.* 2nd ed. Butterworths, London.

LEE, S. and SEO, M. H. (2008). Semiparametric estimation of a binary response model with a change-point due to a covariate threshold. *Journal of Econometrics* **144** 492–499.

LEE, S., SEO, M. H. and SHIN, Y. (2014). The Lasso for high-dimensional regression with a possible change-point. *Journal of the Royal Statistical Society, Series B* Forthcoming.

LOH, P.-L. and WAINWRIGHT, M. J. (2013). Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. In *Advances in Neural Information Processing Systems.*

LOVÁSZ, L. and VEMPALA, S. (2007). The geometry of logconcave functions and sampling algorithms. *Random Structures & Algorithms* **30** 307–358.

MEINSHAUSEN, N. and YU, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *Annals of Statistics* 246–270.

NEGAHBAN, S. N., RAVIKUMAR, P., WAINWRIGHT, M. J. and YU, B. (2012). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science* **27** 538–557.

PONS, O. (2003). Estimation in a Cox regression model with a change-point according to a threshold in a covariate. *Annals of Statistics* **31** 442–463.

RASKUTTI, G., WAINWRIGHT, M. and YU, B. (2010). Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Transactions on Information Theory* **57** 6976–6994.

SEIJO, E. and SEN, B. (2011a). Change-point in stochastic design regression and the bootstrap. *Ann. Statist.* **39** 1580–1607.

SEIJO, E. and SEN, B. (2011b). A continuous mapping theorem for the smallest argmax functional. *Electron. J. Statist.* **5** 421–439.

SUN, T. and ZHANG, C.-H. (2012). Scaled sparse linear regression. *Biometrika* **99** 879–898.

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

TONG, H. (1990). *Non-linear time series: a dynamical system approach.* Oxford University Press.

VAN DE GEER, S. A. (2008). High-dimensional generalized linear models and the lasso. *Annals of Statistics* **36** 614–645.

van de Geer, S. A. and Bühlmann, P. (2009). On the conditions used to prove oracle results for the Lasso. *Electron. J. Stat.* **3** 1360–1392.

van der Vaart, A. and Wellner, J. (1996). *Weak convergence and empirical processes.* Springer, New York.

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

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

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics* 894–942.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67** 301–320.

Zou, H. and Li, R. (2008). One-step sparse estimations in non concave penalized likelihood models. *Annals of Statistics* **36** 1509–1533.