# High Dimensional Models with Unknown Change Point

Sokbae Lee, Yuan Liao, Myung Hwan Seo, and Youngki Shin

**Abstract**

Regression with a change point has been one of the important research areas in econometrics. In variable selection problems of high dimensional models, it is possible that the identities of important variables may change. For instance, the important genes in low temperature may be different from the important genes in high temperature for disease classication. Therefore, the sparsity structure on the regressors should be allowed changable, due to some environmental variable. This makes the variable selection problem more realistic. This paper studies a high-dimensional penalized M-estimation problem with unknown change points. Both inferential theories and selection consistency are obtained.

**Keywords:** structural change, lasso, selection consistency, oracle, scad, quantile

## 1 Introduction

This paper is concerned about estimating high-dimensional models with a possible change-point. To illustrate our estimation problem, consider a probit model in which the distribution of the binary outcome $Y \in \{0,1\}$ depends on a high-dimensional regressor $X \equiv (X_1, ..., X_p)^T$ with a possible change-point $\tau_0$:

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

where $\Phi(t)$ denotes the cumulative distribution function of the standard normal distribution. In addition, there is a scalar variable $Q$ that introduce a possible structural change. Let $\beta_0$ and $\beta_0+\delta_0$ respectively denote the regression coefficients before and after the structural change. The regressors then enter the model through a linear combination $X^T\beta_0 + X^T\delta_0 1\{Q > \tau_0\}$, where $\tau_0 \in \mathcal{T}$ is an unknown threshold parameter, where $\mathcal{T}$ is a bounded interval. Hence $(\beta, \delta, \tau)$ are the unknown parameters.

In this paper, we develop a method for estimating a high-dimensional regression model with a possible change-point due to a covariate threshold, while selecting relevant regressors from a set of many potential covariates. In particular, we propose the $\ell_1$ penalized least squares (Lasso) estimator of parameters, including the unknown threshold parameter, and analyze its properties under a sparsity assumption when the number of possible covariates can be much larger than the sample size.

# 2 High-Dimensional $M$-Estimation with Structural Change

## 2.1 The model

Suppose the outcome variable $Y$ is potentially explained by a high-dimensional vector of regressors $X = (X_1, ..., X_p)^T$, where $p$ is potentially much larger than the sample size $n$. In addition, there is a scalar variable $Q$ that introduce a possible structural change. Let $\beta_0$ and $\beta_0 + \delta_0$ respectively denote the regression coefficients before and after the structural change. The regressors then enter the model through a linear combination $X^T\beta_0 + X^T\delta_0 1\{Q > \tau_0\}$, where $\tau_0 \in \mathcal{T}$ is an unknown threshold parameter, where $\mathcal{T}$ is a bounded interval. Hence $(\beta, \delta, \tau)$ are the unknown parameters. For instance, in the linear quantile regression model

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

where $U$ denotes the regression error. This corresponds to the model $Y = X^T\beta_0 + U$ when $Q \leq \tau_0$ and $Y = X^T(\beta_0 + \delta_0) + U$ when $Q > \tau_0$.

We consider a general $M$-estimation framework, where $(\beta_0, \delta_0, \tau_0)$ are identified as the minimizer of the expected loss:

$$(\beta_0, \delta_0, \tau_0) = \arg\min_{\beta,\delta,\tau} E\rho(Y, X^T\beta + X^T\delta 1\{Q > \tau\}), \tag{2.1}$$

where $\rho(t_1, t_2) : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^+$ is a known convex loss function. For instance, in logistic regression models, $\rho(\cdot, \cdot)$ is the logistic loss. It is assumed that both $\beta_0$ and $\delta_0$ are sparse coefficients so that the number of relevant regressors are small compared to the sample size both before and after the structural change, and only a small portion of the regressors can have structural changes on their effects. In disease classifications based on high-dimensional gene expression data, it is assumd that there are only a small portion of genes contributing to the response variable in both regimes $\{Q > \tau_0\}$ and $\{Q < \tau_0\}$ for an environmental variable $Q$ and certain threshold value $\tau$, and the identifies and effects of relevant genes may differ between the two regimes

Introduce the notation $\alpha = (\beta^T, \delta^T)^T$, $X(\tau) = (X^T, X^T 1\{Q > \tau\})^T$, and $X_i(\tau)$ and $X_{ij}(\tau)$ denote the $i$-th realization of $X(\tau)$ and $j$-th element of $X_i(\tau)$, respectively, $i = 1, ..., n$. We estimate the parameters via the following penalized optimization problem:

$$(\widehat{\alpha}, \widehat{\tau}) = \arg\min_{\alpha,\tau} \frac{1}{n}\sum_{i=1}^{n} \rho(Y_i, X_i(\tau)^T\alpha) + \sum_{j=1}^{2p} \lambda_n |D_j(\tau)\alpha_j|_1 \equiv \arg\min_{\alpha,\tau} S_n(\alpha, \tau), \tag{2.2}$$

where $\lambda_n$ is the tuning parameter and $D_j(\tau)^2 = \frac{1}{n}\sum_{i=1}^{n} X_{ij}(\tau)^2$. The data-dependent weight $D_j(\tau)$ balances regressors adequately. It is worth noting that alternatively, one might penalize $\beta$ and $\beta + \delta$ instead of $\beta$, $\delta$. We opt to penalize $\delta$ in this paper since this formulation makes it convenient to identify the set of regressors whose effects may have structural changes.

From a theoretical standpoint, the threshold model and its estimation (2.1)-(2.2) differ considerably from the regular high-dimensional sparse models with $L_1$-penalizations (e.g., Bickel et al. 2009, van de Geer 2008, Belloni and Chernozhukov 2010). The major difference is that the objective function is neither convex nor smooth in $\tau$, which introduces technical challenges to our theoretical analysis, in particular for nonlinear and possibly nonsmooth loss function $\rho(\cdot,\cdot)$ (as in quantile regression).

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

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

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

**Step 1** For each $k = 1, ..., n$, set $\tau_k = Q_k$. Solve the Lasso problem:

$$\widehat{\alpha}(\tau_k) = \min_{\alpha} S_n(\alpha,\tau_k).$$

**Step 2** Set

$$k^* = \arg\min_{k=1,\ldots,n} S_n(\widehat{\alpha}(\tau_k),\tau_k), \quad \widehat{\alpha} = \widehat{\alpha}(\tau_{k^*}), \quad \widehat{\tau} = \tau_{k^*}$$

In particular, step 2 requires only $n$ function evaluations. If $n$ is very large, $\mathcal{T}_n$ 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 $\widehat{\tau}_N = \arg\min_{\tau\in\mathcal{T}_N} S_n(\widehat{\alpha}(\tau),\tau)$ is a good approximation to $\widehat{\tau}$.

## 2.2 Risk Consistency

Define the *excess risk* to be

$$R(\alpha,\tau) = E\rho(Y, X(\tau)^T\alpha) - E\rho(Y, X(\tau_0)^T\alpha_0).$$

The risk consistency is concerned about the convergence of $R(\hat{\alpha},\widehat{\tau})$. We make the following assumption, which are standard in the literature.

**Assumption 2.1.** *(i) The data $\{(Y_i, X_i, Q_i)\}_{i=1}^n$ are independent and identically distributed with $\mathrm{E}\,|X_{ij}|^m \le \frac{m!}{2}K_1^{m-2}$ for all $j$ and some $K_1 < \infty$ and $\mathrm{E}\left[\left(X^T\delta_0\right)^2 |Q=\tau\right] \le K_2$ for all $\tau \in \mathcal{T}$ and some $K_2 < \infty$.*
*(ii) $\alpha \in \mathcal{A} = \{\alpha : |\alpha|_\infty \le M\}$ for some $M < \infty$, and $\tau \in \mathcal{T} = [\underline{\tau}, \overline{\tau}]$, where the probability of $\{Q > \underline{\tau}\}$ and that of $\{Q < \overline{\tau}\}$ are strictly positive.*
*(iii) The exist universal constants $\underline{D} > 0$ and $\bar{D} > 0$ such that almost surely,*

$$\underline{D} \le \min_{j\le 2p}\inf_{\tau\in\mathcal{T}} D_j(\tau) \le \max_{j\le 2p}\sup_{\tau\in\mathcal{T}} D_j(\tau) \le \bar{D}.$$

*(iv) Let $\mathcal{Y}$ denote the support of $Y$. There is $L > 0$ so that for all $y \in \mathcal{Y}$, and $t_1, t_2 \in \mathbb{R}$,*

$$|\rho(y, t_1) - \rho(y, t_2)| \leq L|t_1 - t_2|$$

Condition (ii) assumes that each regressor of the same magnitude uniformly over the threshold $\tau$. As the data-deponent weights $D_j(\tau)$ are the standard deviations of the regressors, it is not stringent to assume them to be bounded away from both zero and infinity. Condition (iv) requires the Lipchitz continuity of the loss function, which is a standard assumption in the literature for M-estimation (e.g., van der Geer 2009).

Let $J_1$ and $J_2$ denote the nonzero indices of $\beta_0$ and $\delta_0$. Here $J_1$ and $J_2$ can be different, which allows the sets of relevant regressors to be possibly different. Note that $\delta_0$ are the $p+1$ to $2p$ components of $\alpha_0$, so the nonzero indices of $\alpha_0$ belong to

$$J = J_1 \cup \{p + j : j \in J_2\}.$$

For any vector $\alpha \in \mathbb{R}^{2p}$, let $\alpha_J$ and $\alpha_{J^c}$ denote the subvectors whose indies are in $J$ and $J^c$ respectively. The relevant sets before and after the structural change are $\{j : \beta_{0j} \neq 0\}$ and $\{j : \beta_{0j} + \delta_{0j} = 0\}$.

Below, we denote $s = |J|$, as the cardinality of $J$. The following result provides the risk consistency.

**Theorem 2.1** (Risk consistency)**.** *Suppose Assumption 2.1 hold and let*

$$\lambda_n \geq 8\underline{D}^{-1} L \frac{1}{\sqrt{n}} \left( \sqrt{2 \log{(2np)}} + K_1 \log{(2np)} / \sqrt{n} \right) \log_2{(2Mp / |\alpha_0|_1)} \log n. \tag{2.3}$$

*Then,*

$$R(\widehat{\alpha}, \widehat{\tau}) = O_p\left(\lambda_n s\right).$$

The strength of the proposed $l_1$-penalized method is that it is not required to know or pretest whether $\delta_0 = 0$ or not. It is worth noting that we do not have to know whether there exists a threshold $\tau_0$ in the model in order to establish oracle inequalities for the prediction risk and the estimation rates. This implies that we can make prediction and estimation precisely without knowing the presence of threshold.

## 2.3 Threshold Consistency

To begin with, let $\theta_0 = \beta_0 + \delta_0$. Note 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 $\tau_1 < \tau_0$,

$$R(\alpha,\tau_1) = E\left(\rho\left(Y, X^T\beta\right) - \rho\left(Y, X^T\beta_0\right)\right) 1\{Q \le \tau_1\} + E\left(\rho\left(Y, X^T\theta\right) - \rho\left(Y, X^T\theta_0\right)\right) 1\{Q > \tau_0\} \tag{2.4}$$
$$+ E\left(\rho\left(Y, X^T\theta\right) - \rho\left(Y, X^T\beta_0\right)\right) 1\{\tau_1 < Q \le \tau_0\},$$

and when $\tau_2 > \tau_0$,

$$R(\alpha,\tau_2) = E\left(\rho\left(Y, X^T\beta\right) - \rho\left(Y, X^T\beta_0\right)\right) 1\{Q \le \tau_0\} + E\left(\rho\left(Y, X^T\theta\right) - \rho\left(Y, X^T\theta_0\right)\right) 1\{Q > \tau_2\} \tag{2.5}$$
$$+ E\left(\rho\left(Y, X^T\beta\right) - \rho\left(Y, X^T\theta_0\right)\right) 1\{\tau_0 < Q \le \tau_2\}.$$

The consistency of $\widehat{\tau}$ is based on the fact that all the six terms in the above decompositions 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 $R(\widehat{\alpha}, \widehat{\tau})$, which is obtained in Theorem 2.1, and that all the six decomposed terms are non-negative. The non-negativeness holds by the following identification condition.

For any random variable $Z$, write $\|Z\|_2 = (EZ^2)^{1/2}$.

**Assumption 2.2** (Identification of $\alpha_0$). *(i) For all $\alpha \in \mathcal{A}$,*

$$E[\rho(Y, X(\tau_0)^T\alpha) - \rho(Y, X(\tau_0)^T\alpha_0)|Q] \ge 0,$$

*almost surely. Furthermore, the equality holds with probability one if and only if $\alpha = \alpha_0$.*
*(ii) For any $\varepsilon > 0$, there is $C_\varepsilon > 0$*

$$\inf_{\|(X(\tau_0)'(\alpha-\alpha_0))1\{Q\le\tau_0\}\|_2>\varepsilon} E\left(\rho\left(Y, X(\tau_0)^T\alpha\right) - \rho\left(Y, X(\tau_0)^T\alpha_0\right)\right) 1\{Q \le \tau_0\} > C_\varepsilon$$
$$\inf_{\|(X(\tau_0)'(\alpha-\alpha_0))1\{Q>\tau_0\}\|_2>\varepsilon} E\left(\rho\left(Y, X(\tau_0)^T\alpha\right) - \rho\left(Y, X(\tau_0)^T\alpha_0\right)\right) 1\{Q > \tau_0\} > C_\varepsilon.$$

We shall prove in the appendix that under Assumption 2.2 all the six terms on the right hand side of (2.4) and (2.5) are non-negative.

**Remark 2.1.** Assumption 2.2(ii) means that $E\left(\rho\left(Y, X^T\beta\right) - \rho\left(Y, X^T\beta_0\right)\right) 1\{Q \le \tau_0\} \to 0$ and $E\left(\rho\left(Y, X^T\theta\right) - \rho\left(Y, X^T\theta_0\right)\right) 1\{Q > \tau_0\} \to 0$ only if $\left\|\left(X^T(\beta - \beta_0)\right) 1\{Q \le \tau_0\}\right\|_2 \to 0$ and $\left\|\left(X^T(\theta - \theta_0)\right) 1\{Q > \tau_0\}\right\|_2 \to 0$. This condition strengthens the identification of $\alpha_0$. For instance, in the linear model $Y = X(\tau_0)^T\alpha_0 + U$, and $\rho$ is the least squares loss. Assuming $U$ to be independent of $(X, Q)$, then

$$E[\rho(Y, X(\tau_0)^T\alpha) - \rho(Y, X(\tau_0)^T\alpha_0)|Q] = E[(X(\tau_0)^T(\alpha_0 - \alpha))^2|Q].$$

Hence Conditions (i) and (ii) are satisfied with $C_\varepsilon = \varepsilon^2$.

**Remark 2.2.** A sufficient condition for this in terms of conditional expectation?

Furthermore, we impose the following condition to identify $\tau_0$.

**Assumption 2.3** (Identification of $\tau_0$)**.** *For any* $\varepsilon > 0$, $\Pr\{|Q-\tau_0| < \varepsilon\} > 0$ *and*

$$\inf_{\tau\in\mathcal{T}} E\left[\rho\left(Y,\left(X^T, X^T 1\{Q\le\tau_0\}\right)\alpha_0\right) - \rho\left(Y, X(\tau_0)^T\alpha_0\right)\Big| Q=\tau\right] > 0.$$

Recall that $X(\tau_0)^T\alpha_0 = X^T\beta_0 + X^T\delta_0 1\{Q>\tau_0\}$. Hence $X(\tau_0)^T\alpha_0 = X^T\beta_0$ when $Q<\tau_0$ and $X(\tau_0)^T\alpha_0 = X^T\theta_0$ when $Q>\tau_0$. It is the opposite for $\left(X^T, X^T 1\{Q\le\tau_0\}\right)\alpha_0$. This assumption reflects two aspects that determine the identification of $\tau_0$: both $\delta_0$ and the support of $Q$ around $\tau_0$ should be bounded away from zero. For instance, in the linear model, the second condition is satisfied as long as there is $C>0$ so that $E[(X^T(\beta_0-\theta_0))^2|Q] > C$ almost surely. The first condition is satisfied, for example, when $Q$ has a density function that is bounded away from zero in a neighborhood of $\tau_0$.

**Assumption 2.4.** *(moment bounds) There exist* $0 < C_1 \le C_2 < 1$ *such that for all* $\beta\in\mathbb{R}^p$ *such that* $E|X^T\beta| \neq 0$*,*

$$C_1 < \frac{E|X^T\beta|1\{Q>\tau_0\}}{E|X^T\beta|} < C_2.$$

This assumption requires $Q$ has non-negligible supports on both sides of $\tau_0$. Note that it is equivalent to

$$(\frac{1}{C_2}-1)E|X^T\beta|1\{Q>\tau_0\} \le E|X^T\beta|1\{Q<\tau_0\} \le (\frac{1}{C_1}-1)E|X^T\beta|1\{Q>\tau_0\}. \tag{2.6}$$

Hence this assumption also prevents the conditional distribution of $X^T\beta$ given $Q$ from changing too dramatically across regimes.

**Theorem 2.2** (Consistency of $\hat\tau$)**.** *Under Assumptions 2.1-2.4,*

$$\widehat{\tau} \xrightarrow{p} \tau_0.$$

## 2.4 Rate of convergence and super-efficiency

This subsection derives the rate of convergence for $\hat\alpha - \alpha_0$, and proves that we can achieve the super-convergence rate for $\widehat{\tau}-\tau_0$ when $\tau_0$ is identified. The following assumptions are in order.

Recall that when $Q\le\tau_0$, $X(\tau_0)^T\alpha_0 = X^T\beta_0$, while when $Q>\tau_0$, $X(\tau_0)^T\alpha_0 = X^T\theta_0$. Hence we define the "prediction ball" with radius $r$ and corresponding centers as follows:

$$\begin{aligned}\mathcal{B}(\beta_0, r) &= \{\beta\in\mathbb{R}^p : E[(X^T(\beta-\beta_0))^2 1\{Q\le\tau_0\}] \le r^2\}\\ \mathcal{G}(\theta_0, r) &= \{\theta\in\mathbb{R}^p : E[(X^T(\theta-\theta_0))^2 1\{Q>\tau_0\}] \le r^2\}\end{aligned} \tag{2.7}$$

The following assumption is standard in the regression analysis with a change point.

**Assumption 2.5.** *There are $M > 0$, $c_0 > 0$, a neighborhood $\mathcal{T}_0$ of $\tau_0$ and $r > 0$ so that uniformly for $\beta \in \mathcal{B}(\beta_0, r)$ and $\theta \in \mathcal{G}(\theta_0, r)$,*
*(i)* $\inf_{\tau\in\mathcal{T}_0} E[(X^T\delta_0)^2|Q = \tau] > c_0$.
*(ii)* $E[|X^T\beta|^l|Q = \tau]$ *and* $E[|X^T\theta|^l|Q = \tau]$ *are continuous and bounded on* $\tau \in \mathcal{T}_0$, $l = 1, 2$. *In addition,*

$$\sup_{\tau\in\mathcal{T}_0} E(|X^T(\beta - \beta_0)||Q = \tau) < ME|X^T(\beta - \beta_0)|1\{Q > \tau_0\},$$
$$\sup_{\tau\in\mathcal{T}_0} E(|X^T(\theta - \theta_0)||Q = \tau) < ME|X^T(\theta - \theta_0)|1\{Q > \tau_0\}$$

*(iii) $Q$ has a density function that is continuous and bounded away from zero on $\mathcal{T}_0$. Moreover, the conditional distribution of $Q$ given $X$ has a density function $f_{Q|X}(q|x)$ that is bounded uniformly for $q \in \mathcal{T}_0$ and $x$.*

In high-dimensional M-estimations, two conditions are key to the analysis, and are commonly assumed in the literature: the *conditional margin condition* (van de Geer 2009), and the *compatibility condition* (Belloni and Chernozhukov 2011, Bickel et al. 2008, Bulhmann and van de Geer 2010. In particular, the conditional margin condition links the excess risk function to the usual $L_2$ risk. While it can be straightforward to verify in many linear and nonlinear models, we impose its general form as follows:

**Assumption 2.6** (Conditional margin condition)**.** *There exists a constant $c > 0$ so that for all $\tau \in \mathcal{T}_0$,*

$$E\left(\rho\left(Y, X^T\theta_0\right) - \rho\left(Y, X^T\beta_0\right)\right)1\{\tau < Q \le \tau_0\} \ge cE(X^T(\beta_0 - \theta_0))^2 1\{\tau < Q \le \tau_0\}$$
$$E\left(\rho\left(Y, X^T\beta_0\right) - \rho\left(Y, X^T\theta_0\right)\right)1\{\tau_0 < Q \le \tau\} \ge cE(X^T(\beta_0 - \theta_0))^2 1\{\tau_0 < Q \le \tau\},$$

*and there is $r > 0$ so that for all $\beta \in \mathcal{B}(\beta_0, r)$ and $\theta \in \mathcal{G}(\theta_0, r)$,*

$$E\left(\rho\left(Y, X^T\beta\right) - \rho\left(Y, X^T\beta_0\right)\right)1\{Q \le \tau_0\} \ge cE(X^T(\beta - \beta_0))^2 1\{Q \le \tau_0\}$$
$$E\left(\rho\left(Y, X^T\theta\right) - \rho\left(Y, X^T\theta_0\right)\right)1\{Q > \tau_0\} \ge cE(X^T(\theta - \theta_0))^2 1\{Q > \tau_0\}.$$

In the presence of a change point, our margin condition consists of two sets of conditions, one for the regression coefficients and the other for the change point. Combining Assumptions 2.5 and 2.6 imply, for some $c_0 > 0$,

$$E\left(\rho\left(Y, X^T\beta_0\right) - \rho\left(Y, X^T\theta_0\right)\right)1\{\tau_0 < Q \le \tau\} \ge c_0(\tau - \tau_0), \tau > \tau_0 \tag{2.8}$$
$$E\left(\rho\left(Y, X^T\theta_0\right) - \rho\left(Y, X^T\beta_0\right)\right)1\{\tau < Q \le \tau_0\} \ge c_0(\tau_0 - \tau), \tau < \tau_0.$$

Note that in the linear model $\rho(Y, X(\tau)^T\alpha) = (Y - X(\tau)^T\alpha)^2$. Hence in this case it is straightforward to verify Assumption 2.6, and the inequalities can be replaced with equalities with $c = 1$. For nonlinear models, this condition is often satisfied as long as the excess risk can be locally well approximated by quadratic functions (see Bulhmann and van de Geer 2010). Indeed, we shall verify this condition for the binary choice model and quantile regression in Section 4.

The following condition is the well-known *compatibility condition* (see Buhlmann and van de Geer 2011, Chapter 6).

**Assumption 2.7** (Compatibility condition)**.** *There is a constant $\phi(J) > 0$ so that for all $\tau \in \mathcal{T}_0$ and all $\alpha \in \mathbb{R}^{2p}$ satisfying $|\alpha_{J^c}|_1 \leq 5|\alpha_J|_1$,*

$$\phi(J)|\alpha_J|_1^2 \leq s\alpha^T EX(\tau)X(\tau)^T\alpha. \tag{2.9}$$

**Remark 2.3.** This condition is similar to the "restricted eigenvalue condition" as in Bickel et al. (2009), and is commonly assumed in high-dimensional sparse literature, e.g., Candes and Tao (2007), van de Geer et al. (2013), Bickel and Chernozhukov (2011), Belloni et al. (2012), etc. Note that it is *easier* to satisfy than the restricted eigenvalue condition because it is imposed on the population covariance matrix $EX(\tau)X(\tau)^T$. So a simple sufficient condition of Assumption 2.7 is the the smallest eigenvalue of $EX(\tau)X(\tau)^T$ 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 2.1** (Bounded minimum eigenvalue in 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 $\tau$, and $f_i$ is a $k$-dimensional common factors that may not be observable; $u_i$ is a the error term for factor analysis that is independent of $f_i$. Then this is a factor model with a possible structural change on the loading matrix. 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:

$$EX_i(\tau)X_i(\tau)^T = \Lambda(\tau)\mathrm{cov}(f_i)\Lambda(\tau)^T + Eu_iu_i^T.$$

Then a sufficient condition of Assumption 2.7 is that all the eigenvalue of $Eu_iu_i^T$ are bounded below by a constant $c_{\min}$, and (2.9) is satisfied for $\phi(J) = c_{\min}$. Note that assuming the minimum eigenvalue of $Eu_iu_i^T$ to be bounded below is not stringent, because for identifiability purpose, $Eu_iu_i^T$ is often assumed to be diagonal. Then it is sufficient to have $\min_{j\leq 2p} Eu_{ij}^2 > c_{\min}$. □

The following theorem presents the rates of convergence. Define

$$\omega_n = (\log p)(\log n)\sqrt{\frac{\log p}{n}}. \tag{2.10}$$

**Theorem 2.3.** *Suppose $\omega_n |J|^2 \log p = o(1)$. Under Assumptions 2.1- 2.7, there is $C > 0$, for $\lambda_n = C\omega_n$,*

$$|\widehat{\alpha} - \alpha_0|_1 = O_p(\omega_n |J|)$$

$$R(\widehat{\alpha}, \widehat{\tau}) = O_p(\omega_n^2 |J|)$$

*and*

$$|\widehat{\tau} - \tau_0| = O_p(\omega_n^2 |J|).$$

The achieved convergence rate for $\widehat{\alpha}$ is slightly slower than the usual rate for LASSO estimation (e.g., Bickel et al. 2009, Belloni and Chernozhukov 2010), with an additional factor $(\log p)(\log n)$ due to the unknown threshold value $\tau_0$. In the next section we will see that it can be improved to be the oracle rate of convergence after variable selection consistency is achieved via a second-step regularization.

# 3 Oracle Properties and Inferential Theory

## 3.1 SCAD-weighted regularization

This section studies the oracle properties of sparse estimations. The oracle property (see Fan and Li 2001) such as variable selection consistency is one of the fundamental scientific questions for high-dimensional methods. It can enhance the model interpretability with parsimonious representation.

To achieve the variable selection consistency, we further employ a weighted $L_1$ penalized estimator. After the first-step of LASSO regression, we have obtained a consistent sparse estimator $\widehat{\alpha}$ for $\alpha_0$, and an estimator for the change point $\tau$, such that

$$|\widehat{\alpha} - \alpha_0|_1 = O_p(\omega_n s).$$

In addition, we obtain an estimator for the change point $\widehat{\tau}$. When $\tau_0$ is identifiable,

$$|\widehat{\tau} - \tau_0| = O_p(\omega_n^2 s).$$

Otherwise, $\widehat{\tau}$ is some number on the domain of $\tau_0$. Still let $D(\widehat{\tau})$ denote a diagonal matrix of the sample variances of $X(\widehat{\tau})$. These estimators are useful to form the data-dependent weights for the weighted $L_1$ penalty on our second stage for variable selection. Specifically, for some tuning parameter $\mu_n$, and $j = 1, ..., p$ let $w_j$ be the SCAD-weight of Fan and Li (2001), namely,

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

Here $a > 1$ is some prescribed constant, and $a = 3.7$ is often used in the literature.

The objective function is then,

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

We then define the *estimator-of-oracle-property* $\widetilde{\alpha}$ to be the global minimizer of $\widetilde{Q}_n(\alpha)$ on the parameter space of $\alpha$:

$$\widetilde{\alpha} = \arg\min_{\alpha} \widetilde{Q}_n(\alpha).$$

Note that the SCAD penalty is employed here as the weights in the $L_1$ regularization, because the consistent estimation has been achieved on the first stage, which serves as a natural initial values as the "local linear approximation" as in Zou and Li (2006). As a result, the convexity of the objective function is maintained so long as $\rho(y,t)$ is convex, and the estimator can be directly defined as the global minimizer, assumed to be unique. This enhances the computational convenience. In contrast, the regular SCAD penalty is well-known to be folded-concave, whose regularized estimator has often been defined as a local minimizer of the penalized objective function.

### 3.2 Oracle properties

One of the main features of our technicalities is that the loss function $\rho(y,t)$ is allowed to be non-differentiable, and only the continuity is required. On the other hand, it is often the case that the differentiability holds after expectations are taken. Thus models like quantile regressions can be allowed, and this extends the usual GLM (generalized linear model, e.g., Fan and Lv 2011) to a more general M-estimation framework.

We assume $E\rho(Y, X(\tau)^T\alpha)$ be differentiable with respect to $\alpha$, and define

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

Also, let $m_J(\tau,\alpha) = (m_j(\tau,\alpha) : j \in J)$.

**Assumption 3.1.** *$E\rho(Y, X(\tau)^T\alpha)$ is three times differentiable with respect to $\alpha$, and there are $c, c_1, c_2, L > 0$ so that the following conditions hold: for all large $n$ and all $\tau \in (\tau_0 - c, \tau_0 + c)$,*
*(i) there is $M_n > 0$ that may depend on the sample size, so that*

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

*(ii) There is $C > 0$, for all $\alpha$ so that $|\alpha - \alpha_0|_1 < C$,*

$$\max_{j\le 2p} \sup_{|\tau-\tau_0|<c} |m_j(\tau,\alpha) - m_j(\tau,\alpha_0)| < L\,|\alpha - \alpha_0|_1\,.$$

*(iii)* $m(\tau_0, \alpha_0) = 0$, *and*

$$\inf_{|\tau-\tau_0|<c} \lambda_{\min}\left(\frac{\partial^2 E\rho(Y, X_J(\tau)^T\alpha_{0J})}{\partial\alpha_J\partial\alpha_J^T}\right) > c_1.$$

$$\sup_{|\alpha_J-\alpha_{0J}|_1<c_2,} \sup_{|\tau-\tau_0|<c} \max_{i,j,k\in J} \left|\frac{\partial^3 E\rho(Y, X_J(\tau)^T\alpha_J)}{\partial\alpha_i\partial\alpha_j\partial\alpha_k}\right| < L.$$

The score-condition in the population level is expressed by $m(\tau_0, \alpha_0) = 0$ in Condition (iii), which is satisfied by most of the statistical models for M-estimations, including, e.g., generalized linear models and quantile regressions. Conditions (i) and (ii) regulate the continuity of the score $m(\tau, \alpha)$ and the loss $\rho(y, t)$. Condition (i) requires the Lipschitz continuity of the score function with respect to the threshold. The Lipschitz constant may grow with $n$, due to it is assumed uniformly over $j \le 2p$. In many interesting examples being considered, $M_n$ in fact grows very slowly so does not affect the asymptotic behavior of the estimators. In the examples being considered (the binary and quantile regression models), we will show that $M_n = Ms^{3/2}$ for some $M > 0$. Condition (ii) requires the equi-continuity at $\alpha_0$ in the $L_1$-norm of the class

$$\{m_j(\tau, \alpha) : \tau \in (\tau_0 - c, \tau_0 + c), j \le 2p\}.$$

In Section 4, we shall verify the above assumption in both the binary regression and quantile regression models.

Under the foregoing assumptions, the following two theorems establish the oracle properties of the weighted-$L_1$-regularized estimators.

**Theorem 3.1.** *Suppose $s^4(\log s) = o(n)$, the foregoing assumptions are satisfied, and*

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

*If we partition $\widetilde{\alpha} = (\widetilde{\alpha}_J, \widetilde{\alpha}_{J^c})$ so that $\widetilde{\alpha}_J = (\widetilde{\alpha}_j : j \in J)$ and $\widetilde{\alpha}_{J^c} = (\widetilde{\alpha}_j : j \notin J)$, then*

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

*and*

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

This theorem covers both the cases of $\delta_0 = 0$ and $\delta_0 \ne 0$. In particular, when $\delta_0 = 0$, $\tau_0$ is not identifiable, and hence $\alpha_{0J} = \beta_{0J}$. This theorem then implies $\widehat{\delta} = 0$ with probability approaching one. Also in this case, Conditions (i) in Assumption 3.1 is trivially satisfied because $m(\tau, \alpha_0)$ is free of $\tau$.

Some additional remarks are in order.

**Remark 3.1.** Because $\min_{j\in J} |\widetilde{\alpha}_j| \ge \min_{j\in J} |\alpha_{0J}| - |\widetilde{\alpha}_J - \alpha_{0J}|_2$, hence $\min_{j\in J} |\widetilde{\alpha}_j| > 0$ with probability approaching one. If we further define $\widehat{J} = \{j : \widetilde{\alpha}_j \ne 0\}$, then $P(\widehat{\alpha}_{J^c} = 0) \to 1$ implies the variable selection consistency:

$$P(\widehat{J}=J)\to 1.$$

In addition, the required strength of $\nu_n$ is stronger than that of the tuning parameter $\lambda_n$ for the first-stage estimation. The choice of tuning parameters depends on the purpose of estimation, and indeed it is known that the variable selection consistency often requires a larger tuning parameter than does prediction (e.g., Sun and Zhang 2013).

**Remark 3.2.** We have achieved the oracle rate of convergence $O_p(\sqrt{s\log s/n})$ in the $L_2$-distance. Compared to the convergence rate in Theorem ??, we see that after the consistent variable selection, the rate of convergence can be improved, and is slightly faster than the sparse minimax rate $O_p(\sqrt{s\log p/n})$ in, e.g., Johnstone (1994) and Raskutti et al. (2011). It is natural that the oracle rate is faster than the minimax rate. Intuitively, when the variable selection consistency is established, we have recovered the true set of relevant regressors with a high probability. Then estimation can be restricted on this oracle set, which becomes a classical low-dimensional problem. As a result, the price $\sqrt{\log p}$ for not knowing $J$ can be avoided.

**Theorem 3.2** (Asymptotic Normality)**.** *Under the conditions of Theorem 3.1, for any unit vector $v\in\mathbb{R}^s$, $|v|_2=1$,*

$$\sqrt{n}v^T\Sigma_1^{-1/2}\Sigma_2(\widetilde{\alpha}_J-\alpha_{0J})\to^d N(0,1)$$

*where $\Sigma_1=\mathrm{var}(\frac{\partial}{\partial\alpha_J}\rho(Y_i,X_{iJ}^T(\tau_0)\alpha_{0J}))$ and $\Sigma_2=E\frac{\partial^2}{\partial\alpha_J\partial\alpha_J^T}\rho(Y_i,X_{iJ}^T(\tau_0)\alpha_{0J})$.*

**Remark 3.3.** The asymptotic normality presented here reviews 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 $J$, and (2) the effect of estimating $\tau_0$ is negligible when $\tau_0$ is identifiable. Hence the limiting distribution is also the same as if $\tau_0$ 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 $\tau_0$. Consequently, there is no efficiency loss.

# 4 Applications to Quantile Regression and Logistic Models

Assumptions 2.6 and 3.1 are key assumptions of the loss function $\rho(.,.)$ and the model being considered. We give two interesting examples that fall into our context, where low-level conditions are presented to verify our regularity assumptions.

## 4.1 Quantile regression with a change point

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

$$\begin{aligned} Y &= X^T\beta_0+X^T\delta_0 1\{Q>\tau_0\}+U, \\ &= X(\tau)^T\alpha_0+U, \qquad P(U\le 0|X,Q)=\gamma \end{aligned}$$

for some known $\gamma \in (0,1)$. The high-dimensional quantile regression has caused attentions in the recent years. Belloni and Chernozhukov (2011) studied $L_1$-regularized estimation for $\beta_0$ without a change point ($\delta_0 = 0$). Fan et al. (2013) studied the variable selection consistency when $U$ has a heavy-tailed distribution.

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 $\alpha$. But it is straightforward to check that in this case

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

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

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

Because $E[X_j(\tau_0)(1\{Y - X(\tau_0)^T\alpha_0 \leq 0\}] = E[X_j(\tau_0)P(U \leq 0|X,Q)]$. Therefore $m_j(\tau_0, \alpha_0) = 0$ for all $j = 1, ..., 2p$. In addition, we assume $\Gamma(\tau, \alpha_0)$ to be positive definite uniformly in a neighborhood of $\tau_0$.

A sufficient condition for Assumptions 2.6 and 3.1 is given as follows:

**Assumption 4.1.** *(i) The conditional distribution $Y|X,Q$ has a differentiable density function $f_{Y|X,Q}(y|x,q)$, whose derivative with respect to $y$ is denoted by $f'_{Y|X,Q}(y|x,q)$. There are constants $C_1, C_2, C_3 > 0$ so that for all $(y,x,q)$ in the support of $(Y,X,Q)$,*

$$|f'_{Y|X,Q}(y|x,q)| < C_1, \quad C_2 < f_{Y|X,Q}(y|x,q) < C_3.$$

*(ii) There are $c > 0, L > 0$, for all $\tau_1, \tau_2 \in (\tau_0 - c, \tau_0 + c)$, $P(\tau_1 < Q < \tau_2) < L|\tau_1 - \tau_2|$. $\sup_{q\in\mathcal{Q}} \max_{j\leq p} E(X_j^2|Q = q) < C_3$.*
*(iii) For $C_1, C_2$ as in (i), there is $r > 0$, for all $\beta \in \mathcal{B}(\beta_0, r)$ and $\delta \in \mathcal{G}(\beta_0, r)$ (as defined in (2.7)), almost surely,*

$$\frac{3C_2}{2C_1}\frac{E[(X^T\delta_0)^2|Q]}{E[|X^T\delta_0|^3|Q]} \geq 1, \quad \frac{3C_2}{2C_1}\frac{E[(X^T(\beta-\beta_0))^2|Q]}{E[|X^T(\beta-\beta_0)|^3|Q]} \geq 1, \quad \frac{3C_2}{2C_1}\frac{E[(X^T(\theta-\theta_0))^2|Q]}{E[|X^T(\theta-\theta_0)|^3|Q]} \geq 1.$$

A sufficient condition of Condition (iii) is known as the "restricted nonlinearity", as in condition D.4 in Belloni and Chernozhukov (2011).

The following result verifies Assumptions 2.6 and 3.1.

**Lemma 4.1.** *Suppose Assumption 4.1 holds, then Assumptions 2.6 and 3.1 hold, with $M_n = Ms$ for some $M > 0$.*

As a result, a straightforward application of Theorems 3.1 implies the variable selection consistency. As for the limiting distribution, we have:

## 4.2 Binary regression models 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 $g(t) : \mathbb{R} \to (0,1)$ is a known differentiable function. Typical examples of $g(t)$ are:

$$\textbf{Logitmodel}: \quad g(t) = \frac{e^t}{1+e^t}$$

$$\textbf{Probitmodel}: \quad g(t) = \Phi(t)$$

where $\Phi(t)$ denotes the cumulative distribution function of the standard normal distribution. 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 $g'(t)$ denotes the derivative of $g(t)$. Immediately, $m_j(\tau_0,\alpha_0) = 0$ for $j = 1, ..., 2p$.

A sufficient condition for Assumptions 2.6 and 3.1 is given as follows:

**Assumption 4.2.** *There are $C_1, C_2, r, L > 0$ and $\epsilon > 0$ so that for all $\tau, \tau_1, \tau_2 \in (\tau_0 - r, \tau_0 + r)$,*
*(i) $g(t) : \mathbb{R} \to (0,1)$ twice differentiable and* $\sup_{t\in\mathbb{R}} |g'(t)| < C_1, \sup_{t\in\mathbb{R}} |g''(t)| < C_1$.
*(ii) Almost surely $\epsilon < g(X(\tau)^T\alpha) < 1-\epsilon$ for all $|\alpha - \alpha_0|_1 < r$.*
*(iii) $P(\tau_1 < Q < \tau_2) < L|\tau_1 - \tau_2|$.* $\sup_{q\in\mathcal{Q}} \max_{j\le p} E(X_j^2|Q=q) < C_1$.
*(iv) $\log g(t)$ and $\log(1-g(t))$ are Lipschitz continuous in $t$.*

In particular, Condition (ii) requires $g(X(\tau)^T\alpha)$ be bounded away from both zero and one in an $L_1$-neighborhood of $\alpha_0$. Intuitively, we should have observations for both $Y = 1$ and $Y = 0$ almost everywhere within the support of $(X, Q)$. When Condition (ii) is violated, however, often we can still proceed inferences. For instance, for binary classification purposes, we only classify the new outcomes in the region of $(X, Q)$ so that there are sufficient observations for both $Y = 1$ and $Y = 0$, because the classification in the region where one of the outcomes is missing from the dataset is relatively straightforward. In addition, the information that $P(Y = 1|X, Q) = 0$ or 1 for particular $(X, Q)$ can often be transferred to refine the parameter space of $\alpha_0$. Information of this kind also helps pre-screening out irrelevant variables. Nevertheless, we can always focus on the regions of $(X, Q)$ so that Condition (ii) is satisfied.

The following result verifies Assumptions 2.6 and 3.1.

**Lemma 4.2.** *Suppose Assumption 4.2 holds, then Assumptions 2.6 and 3.1 hold, with $M_n = Ms$ for some $M > 0$.*

# 5 Monte Carlo Experiments

In this section we provide the results of some Monte Carlo simulation studies. We consider two baseline models: the linear regression and the logistic regression. Thus, each model can be written as

$$\begin{aligned} Y_i &= X_i'\beta_0 + 1\{Q_i < \tau_0\}X_i'\delta_0 + \varepsilon_{i1} \\ Y_i &= 1\{X_i'\beta_0 + 1\{Q_i < \tau_0\}X_i'\delta_0 + \varepsilon_{i2} > 0\} \end{aligned}$$

where $\varepsilon_{i1}$ is generated from the standard normal distribution, and $\varepsilon_{i2}$ is generated from the logistic distribution. In both models, $X_i$ is a $p$-dimensional vector generated from $N(0, I)$, $Q_i$ is a scalar generated from the uniform distribution on the interval of $(0, 1)$. The $(p \times 1)$ parameters $\beta_0$ and $\delta_0$ are set to $\beta_0 = (1, 0, 1, 0, \ldots, 0)$ and $\delta = (0, 1, 1, 0, \ldots, 0)$, respectively, and the threshold parameter $\tau_0$ is set to 0.5. The sample size is set to 400. The different size of $X_i$ are considered as $p = 50, 100, 200$, and 400, so that the total number of regressors are $2p = 100, 200, 400$ and 800, respectively. The range of $\tau$ is set to $\mathbb{T} = [0.15, 0.85]$. We conduct 1,000 replications of each design.

We can estimate the model by the standard algorithm for the lasso estimator such as LARS by Efton et al (2004) or GLMNET by [reference ()] without much modification. In each step, given all tuning parameters, we estimate the model for each grid point of $\tau$ spanning over 71 equi-spced points on $\mathbb{T}$. This procedure is identical to the standard linear lasso. Next, we choose $\widehat{\tau}$ and corresponding $\widehat{\alpha}(\tau)$ that minimize the profiled objective function. We set the tuning parameters in each step as

$$\lambda_1 = A_1\sqrt{\frac{\log 2p}{n}} \quad \text{and} \quad \lambda_2 = A_2\sqrt{\frac{\log 2p}{n}},$$

where the constants $A_1$ and $A_2$ vary as $A_1 = \{0.05, 0.10, \ldots, 0.50\}$ and $A_2 = \{0.5, 1.0\}$. We also set $a = 3.7$ in the second step SCAD estimator following the convention.

Tables 1–8 summarize these simulation results. In the top panel of each table, we report the results from the first step lasso estimation with different values of $A_1$. In the middle and the bottom panels, we report the results from the second step SCAD estimation with $A_2 = 0.5$ and $A_2 = 1$, respectively. To compare the result, we report mean and median excess risks, the average number of nonzero parameter estimates, the probability that the true nonzero parameters are selected. We also report $\ell_1$ errors of $\widehat{\alpha}$ and $\widehat{\tau}$.

Overall, the results are satisfactory and provide finite sample evidence for the theoretical results we develop in the previous sections. First, we focus on Tables 1–4 of the linear regression model. In all simulation designs, the proposed two-step estimation (SCAD) decreases Excess Risk substantially regardless of the size of $A_2$. Furthermore, the sparse model structure is well-captured when $A_2 = 1.0$. We can also confirm that $\ell_1$ error of $\widehat{\alpha}$ is much smaller in SCAD. It is noteworthy that the latter two results seem to be uniform across different values of the tuning parameter of the first stage LASSO.

We next turn our attention to Tables 5–8 of the logistic regression. Overall, the pattern of the results are very similar to those of the linear regression. The proposed estimation shows good performance in model selection although the improvement from the first step is not as large as the previous linear regression case. However, this seems to be natural as the estimation problem of the logistic regression is more difficult than that of the linear regression. Since we miss the true nonzero parameters in some simulations, we provide separate probabilities such that each nonzero coefficients $(\beta_1, \beta_3, \delta_2, \delta_2)$ are included. Among those values of tuning parameters, the proposed estimator works well unless they are too small (e.g. $A_1 = 0.05$ and $A_2 = 0.5$) or too large, e.g. $(A_1 = 0.50$ and $A_2 = 1)$.

In sum, the proposed two step estimation procedure works well in finite samples and confirms the theoretical results developed earlier. Therefore, it will be useful in the class of high-dimensional threshold regression models, where the estimator can be defined as a minimizer of a convex loss function.

# A Proofs for Section 2

Throughout the proof, we define the empirical process

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

And without loss of generality let $v_n(\alpha_J,\tau)$ indicate $n^{-1}\sum_{i=1}^{n}\left[\rho\left(Y_i, X_{iJ}(\tau)^T\alpha_J\right) - E\rho\left(Y_i, X_{iJ}(\tau)^T\alpha_J\right)\right].$ Also define $D(\tau) = \operatorname{diag}(D_j(\tau) : j \le 2p)$ and then $D_0 = D(\tau_0)$ and $\widehat{D} = D(\widehat{\tau})$. The following lemma bounds the empirical process.

Fix some $K_1 > 0$, define

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

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

$$\Pr\left\{\sup_{\tau\in\mathcal{T}}\sup_{|\alpha-\alpha_0|_1\le m_{1n}}|\nu_n(\alpha,\tau) - \nu_n(\alpha_0,\tau)| \ge a_n m_{1n}\right\} \le \delta, \tag{A.1}$$

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

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

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

**Proof of** (A.1): Let $\{\eta_i\}$ be an iid Rademacher sequence, independent of $\{X_i\}$. By the symmetrization theorem, and for $k \ge \log_2(m_{2n}/m_{1n})$, and then by the contraction theorem,

$$\begin{aligned}
&\mathrm{E}\sup_{\tau}\sup_{|\alpha-\alpha_0|_1\le m_{1n}}|\nu_n(\alpha,\tau) - \nu_n(\alpha_0,\tau)| \\
&\le 2\mathrm{E}\left(\sup_{\tau}\sup_{|\alpha-\alpha_0|_1\le m_{1n}}\left|\frac{1}{n}\sum_{i=1}^{n}\eta_i\rho\left(Y_i, X_i(\tau)^T\alpha\right) - \rho\left(Y_i, X_i(\tau)^T\alpha_0\right)\right|\right) \\
&\le 2L\mathrm{E}\left(\sup_{\tau}\sup_{|\alpha-\alpha_0|_1\le m_{1n}}\left|\frac{1}{n}\sum_{i=1}^{n}\eta_i X_i(\tau)^T(\alpha-\alpha_0)\right|\right).
\end{aligned}$$

Table 1: Linear Regression: $n = 400$, $p = 50$

| Constant | Constant | Excess Risk | | $E[\widehat{J}]\left(E[\widehat{J}_1]/E[\widehat{J}_2]\right)$ | $P\{J_0 \subset \widehat{J}\}$ | $E\left\|\widehat{\alpha} - \alpha_0\right\|_1$ (on $J/J^c$) | $E\left\|\widehat{\tau} - \tau_0\right\|_1$ |
|---|---|---|---|---|---|---|---|
| for $\lambda_1$ | for $\lambda_2$ | Mean | Median | | | | |
| LASSO | | | | | | | |
| 0.05 | NA | 0.046 | 0.048 | 86.90 ( 43.5 / 43.5 ) | 1 | 5.746 ( 0.298 / 5.474 ) | 0.000 |
| 0.10 | NA | 0.059 | 0.058 | 75.09 ( 37.6 / 37.5 ) | 1 | 4.287 ( 0.289 / 4.044 ) | 0.001 |
| 0.15 | NA | 0.072 | 0.071 | 64.92 ( 32.6 / 32.3 ) | 1 | 3.292 ( 0.286 / 3.069 ) | 0.000 |
| 0.20 | NA | 0.087 | 0.085 | 56.01 ( 28.3 / 27.7 ) | 1 | 2.600 ( 0.288 / 2.391 ) | 0.001 |
| 0.25 | NA | 0.101 | 0.100 | 48.52 ( 24.5 / 24.0 ) | 1 | 2.110 ( 0.292 / 1.912 ) | 0.000 |
| 0.30 | NA | 0.116 | 0.113 | 42.14 ( 21.4 / 20.7 ) | 1 | 1.754 ( 0.298 / 1.565 ) | 0.000 |
| 0.35 | NA | 0.130 | 0.128 | 36.66 ( 18.7 / 18.0 ) | 1 | 1.489 ( 0.306 / 1.308 ) | 0.000 |
| 0.40 | NA | 0.145 | 0.144 | 32.15 ( 16.4 / 15.7 ) | 1 | 1.285 ( 0.314 / 1.110 ) | 0.000 |
| 0.45 | NA | 0.159 | 0.158 | 28.21 ( 14.4 / 13.8 ) | 1 | 1.123 ( 0.324 / 0.954 ) | 0.000 |
| 0.50 | NA | 0.174 | 0.172 | 24.51 ( 12.5 / 12.0 ) | 1 | 0.994 ( 0.334 / 0.829 ) | 0.000 |
| SCAD | | | | | | | |
| 0.05 | 0.5 | 0.011 | 0.011 | 24.65 ( 11.8 / 12.8 ) | 1 | 1.186 ( 0.251 / 0.998 ) | 0.000 |
| 0.10 | 0.5 | 0.010 | 0.011 | 23.75 ( 11.6 / 12.1 ) | 1 | 1.101 ( 0.250 / 0.914 ) | 0.001 |
| 0.15 | 0.5 | 0.009 | 0.010 | 23.15 ( 11.5 / 11.7 ) | 1 | 1.032 ( 0.249 / 0.847 ) | 0.000 |
| 0.20 | 0.5 | 0.009 | 0.011 | 22.75 ( 11.4 / 11.3 ) | 1 | 0.978 ( 0.248 / 0.794 ) | 0.001 |
| 0.25 | 0.5 | 0.008 | 0.009 | 22.58 ( 11.5 / 11.1 ) | 1 | 0.937 ( 0.248 / 0.754 ) | 0.000 |
| 0.30 | 0.5 | 0.008 | 0.009 | 22.48 ( 11.4 / 11.0 ) | 1 | 0.905 ( 0.247 / 0.723 ) | 0.000 |
| 0.35 | 0.5 | 0.007 | 0.008 | 22.45 ( 11.5 / 11.0 ) | 1 | 0.882 ( 0.246 / 0.700 ) | 0.000 |
| 0.40 | 0.5 | 0.007 | 0.009 | 22.46 ( 11.5 / 11.0 ) | 1 | 0.866 ( 0.246 / 0.685 ) | 0.000 |
| 0.45 | 0.5 | 0.007 | 0.008 | 22.46 ( 11.5 / 11.0 ) | 1 | 0.853 ( 0.245 / 0.672 ) | 0.000 |
| 0.50 | 0.5 | 0.006 | 0.009 | 22.47 ( 11.5 / 11.0 ) | 1 | 0.844 ( 0.245 / 0.664 ) | 0.000 |
| | | | | | | | |
| 0.05 | 1.0 | 0.008 | 0.008 | 6.95 ( 3.2 / 3.7 ) | 1 | 0.340 ( 0.240 / 0.169 ) | 0.000 |
| 0.10 | 1.0 | 0.007 | 0.007 | 6.63 ( 3.1 / 3.5 ) | 1 | 0.323 ( 0.240 / 0.152 ) | 0.001 |
| 0.15 | 1.0 | 0.006 | 0.006 | 6.45 ( 3.1 / 3.4 ) | 1 | 0.312 ( 0.239 / 0.141 ) | 0.000 |
| 0.20 | 1.0 | 0.006 | 0.007 | 6.33 ( 3.1 / 3.3 ) | 1 | 0.304 ( 0.239 / 0.134 ) | 0.001 |
| 0.25 | 1.0 | 0.005 | 0.005 | 6.23 ( 3.1 / 3.2 ) | 1 | 0.298 ( 0.239 / 0.128 ) | 0.000 |
| 0.30 | 1.0 | 0.005 | 0.004 | 6.16 ( 3.1 / 3.1 ) | 1 | 0.294 ( 0.238 / 0.124 ) | 0.000 |
| 0.35 | 1.0 | 0.005 | 0.004 | 6.13 ( 3.1 / 3.1 ) | 1 | 0.291 ( 0.238 / 0.121 ) | 0.000 |
| 0.40 | 1.0 | 0.005 | 0.005 | 6.12 ( 3.1 / 3.1 ) | 1 | 0.288 ( 0.238 / 0.119 ) | 0.000 |
| 0.45 | 1.0 | 0.004 | 0.006 | 6.10 ( 3.1 / 3.0 ) | 1 | 0.286 ( 0.237 / 0.117 ) | 0.000 |
| 0.50 | 1.0 | 0.004 | 0.006 | 6.10 ( 3.1 / 3.0 ) | 1 | 0.285 ( 0.237 / 0.116 ) | 0.000 |

Table 2: Linear Regression: $n = 400$, $p = 100$

| Constant | Constant | Excess Risk | | $E[\widehat{J}]\left(E[\widehat{J}_1]/E[\widehat{J}_2]\right)$ | $P\{J_0 \subset \widehat{J}\}$ | $E\left\|\widehat{\alpha} - \alpha_0\right\|_1$ (on $J/J^c$) | $E\left\|\widehat{\tau} - \tau_0\right\|_1$ |
|---|---|---|---|---|---|---|---|
| for $\lambda_1$ | for $\lambda_2$ | Mean | Median | | | | |
| LASSO | | | | | | | |
| 0.05 | NA | 0.083 | 0.084 | 166.50 ( 83.5 / 83.0 ) | 1 | 11.840 ( 0.342 / 11.549 ) | 0.002 |
| 0.10 | NA | 0.091 | 0.091 | 138.60 ( 69.9 / 68.7 ) | 1 | 8.042 ( 0.318 / 7.796 ) | 0.003 |
| 0.15 | NA | 0.101 | 0.101 | 116.44 ( 59.1 / 57.4 ) | 1 | 5.824 ( 0.311 / 5.600 ) | 0.002 |
| 0.20 | NA | 0.114 | 0.113 | 98.27 ( 50.3 / 48.0 ) | 1 | 4.415 ( 0.311 / 4.207 ) | 0.001 |
| 0.25 | NA | 0.128 | 0.126 | 83.54 ( 43.1 / 40.5 ) | 1 | 3.470 ( 0.314 / 3.275 ) | 0.001 |
| 0.30 | NA | 0.142 | 0.139 | 71.32 ( 37.0 / 34.3 ) | 1 | 2.800 ( 0.320 / 2.613 ) | 0.001 |
| 0.35 | NA | 0.156 | 0.152 | 61.16 ( 31.9 / 29.3 ) | 1 | 2.307 ( 0.327 / 2.127 ) | 0.001 |
| 0.40 | NA | 0.170 | 0.167 | 52.50 ( 27.4 / 25.1 ) | 1 | 1.932 ( 0.336 / 1.758 ) | 0.001 |
| 0.45 | NA | 0.185 | 0.182 | 45.03 ( 23.5 / 21.6 ) | 1 | 1.637 ( 0.346 / 1.469 ) | 0.001 |
| 0.50 | NA | 0.199 | 0.197 | 38.60 ( 20.0 / 18.6 ) | 1 | 1.402 ( 0.356 / 1.239 ) | 0.000 |
| SCAD | | | | | | | |
| 0.05 | 0.5 | 0.020 | 0.019 | 41.12 ( 19.5 / 21.6 ) | 1 | 1.967 ( 0.264 / 1.769 ) | 0.002 |
| 0.10 | 0.5 | 0.018 | 0.018 | 38.91 ( 19.1 / 19.9 ) | 1 | 1.728 ( 0.261 / 1.536 ) | 0.003 |
| 0.15 | 0.5 | 0.016 | 0.015 | 37.58 ( 18.8 / 18.8 ) | 1 | 1.564 ( 0.258 / 1.373 ) | 0.002 |
| 0.20 | 0.5 | 0.014 | 0.013 | 36.69 ( 18.7 / 18.0 ) | 1 | 1.453 ( 0.257 / 1.264 ) | 0.001 |
| 0.25 | 0.5 | 0.013 | 0.012 | 36.31 ( 18.7 / 17.6 ) | 1 | 1.378 ( 0.255 / 1.190 ) | 0.001 |
| 0.30 | 0.5 | 0.013 | 0.011 | 36.22 ( 18.8 / 17.4 ) | 1 | 1.326 ( 0.254 / 1.139 ) | 0.001 |
| 0.35 | 0.5 | 0.011 | 0.009 | 36.25 ( 18.8 / 17.4 ) | 1 | 1.288 ( 0.253 / 1.102 ) | 0.001 |
| 0.40 | 0.5 | 0.011 | 0.009 | 36.31 ( 18.9 / 17.4 ) | 1 | 1.264 ( 0.252 / 1.078 ) | 0.001 |
| 0.45 | 0.5 | 0.010 | 0.008 | 36.33 ( 18.9 / 17.4 ) | 1 | 1.246 ( 0.252 / 1.061 ) | 0.001 |
| 0.50 | 0.5 | 0.010 | 0.008 | 36.34 ( 18.9 / 17.4 ) | 1 | 1.234 ( 0.252 / 1.049 ) | 0.000 |
| | | | | | | | |
| 0.05 | 1.0 | 0.015 | 0.017 | 8.73 ( 3.8 / 4.9 ) | 1 | 0.400 ( 0.249 / 0.223 ) | 0.002 |
| 0.10 | 1.0 | 0.013 | 0.017 | 8.01 ( 3.7 / 4.3 ) | 1 | 0.359 ( 0.247 / 0.183 ) | 0.003 |
| 0.15 | 1.0 | 0.011 | 0.015 | 7.57 ( 3.6 / 4.0 ) | 1 | 0.339 ( 0.247 / 0.163 ) | 0.002 |
| 0.20 | 1.0 | 0.010 | 0.013 | 7.36 ( 3.6 / 3.8 ) | 1 | 0.328 ( 0.246 / 0.152 ) | 0.001 |
| 0.25 | 1.0 | 0.009 | 0.012 | 7.23 ( 3.5 / 3.7 ) | 1 | 0.322 ( 0.246 / 0.146 ) | 0.001 |
| 0.30 | 1.0 | 0.009 | 0.010 | 7.16 ( 3.5 / 3.6 ) | 1 | 0.317 ( 0.245 / 0.142 ) | 0.001 |
| 0.35 | 1.0 | 0.008 | 0.010 | 7.13 ( 3.5 / 3.6 ) | 1 | 0.314 ( 0.245 / 0.138 ) | 0.001 |
| 0.40 | 1.0 | 0.008 | 0.010 | 7.12 ( 3.5 / 3.6 ) | 1 | 0.312 ( 0.244 / 0.136 ) | 0.001 |
| 0.45 | 1.0 | 0.007 | 0.009 | 7.11 ( 3.5 / 3.6 ) | 1 | 0.310 ( 0.244 / 0.135 ) | 0.001 |
| 0.50 | 1.0 | 0.007 | 0.009 | 7.11 ( 3.5 / 3.6 ) | 1 | 0.309 ( 0.244 / 0.133 ) | 0.000 |

Table 3: Linear Regression: $n = 400$, $p = 200$

| Constant | Constant | Excess Risk | | $E[\widehat{J}]\left(E[\widehat{J}_1]/E[\widehat{J}_2]\right)$ | $P\{J_0 \subset \widehat{J}\}$ | $E\left\|\widehat{\alpha} - \alpha_0\right\|_1$ (on $J/J^c$) | $E\left\|\widehat{\tau} - \tau_0\right\|_1$ |
|---|---|---|---|---|---|---|---|
| for $\lambda_1$ | for $\lambda_2$ | Mean | Median | | | | |
| LASSO | | | | | | | |
| 0.05 | NA | 0.187 | 0.188 | 286.12 ( 147.5 / 138.6 ) | 1 | 21.527 ( 0.410 / 21.256 ) | 0.001 |
| 0.10 | NA | 0.160 | 0.161 | 228.70 ( 120.4 / 108.3 ) | 1 | 13.057 ( 0.355 / 12.838 ) | 0.001 |
| 0.15 | NA | 0.156 | 0.153 | 188.60 ( 100.8 / 87.8 ) | 1 | 9.077 ( 0.339 / 8.878 ) | 0.001 |
| 0.20 | NA | 0.160 | 0.156 | 157.83 ( 85.2 / 72.6 ) | 1 | 6.741 ( 0.333 / 6.556 ) | 0.001 |
| 0.25 | NA | 0.168 | 0.167 | 132.83 ( 72.1 / 60.8 ) | 1 | 5.222 ( 0.336 / 5.045 ) | 0.000 |
| 0.30 | NA | 0.178 | 0.176 | 112.55 ( 61.2 / 51.3 ) | 1 | 4.155 ( 0.340 / 3.986 ) | 0.000 |
| 0.35 | NA | 0.189 | 0.186 | 95.63 ( 52.0 / 43.6 ) | 1 | 3.372 ( 0.346 / 3.208 ) | 0.000 |
| 0.40 | NA | 0.200 | 0.196 | 81.17 ( 44.0 / 37.2 ) | 1 | 2.770 ( 0.354 / 2.611 ) | 0.000 |
| 0.45 | NA | 0.212 | 0.208 | 68.87 ( 37.1 / 31.8 ) | 1 | 2.302 ( 0.363 / 2.146 ) | 0.001 |
| 0.50 | NA | 0.224 | 0.220 | 58.21 ( 31.2 / 27.0 ) | 1 | 1.928 ( 0.374 / 1.775 ) | 0.000 |
| SCAD | | | | | | | |
| 0.05 | 0.5 | 0.038 | 0.040 | 65.29 ( 31.2 / 34.1 ) | 1 | 3.099 ( 0.270 / 2.906 ) | 0.001 |
| 0.10 | 0.5 | 0.029 | 0.027 | 60.16 ( 30.5 / 29.6 ) | 1 | 2.512 ( 0.256 / 2.331 ) | 0.001 |
| 0.15 | 0.5 | 0.023 | 0.020 | 57.53 ( 30.0 / 27.5 ) | 1 | 2.215 ( 0.249 / 2.039 ) | 0.001 |
| 0.20 | 0.5 | 0.019 | 0.018 | 56.21 ( 29.8 / 26.4 ) | 1 | 2.036 ( 0.245 / 1.862 ) | 0.001 |
| 0.25 | 0.5 | 0.017 | 0.017 | 55.51 ( 29.8 / 25.8 ) | 1 | 1.927 ( 0.244 / 1.753 ) | 0.000 |
| 0.30 | 0.5 | 0.015 | 0.015 | 55.32 ( 29.8 / 25.5 ) | 1 | 1.850 ( 0.243 / 1.677 ) | 0.000 |
| 0.35 | 0.5 | 0.014 | 0.013 | 55.34 ( 29.9 / 25.5 ) | 1 | 1.800 ( 0.242 / 1.627 ) | 0.000 |
| 0.40 | 0.5 | 0.013 | 0.011 | 55.44 ( 29.9 / 25.5 ) | 1 | 1.764 ( 0.241 / 1.593 ) | 0.000 |
| 0.45 | 0.5 | 0.012 | 0.010 | 55.53 ( 30.0 / 25.5 ) | 1 | 1.740 ( 0.240 / 1.569 ) | 0.001 |
| 0.50 | 0.5 | 0.011 | 0.008 | 55.51 ( 30.0 / 25.5 ) | 1 | 1.722 ( 0.240 / 1.551 ) | 0.000 |
| 0.05 | 1.0 | 0.025 | 0.022 | 11.07 ( 4.6 / 6.4 ) | 1 | 0.482 ( 0.247 / 0.309 ) | 0.001 |
| 0.10 | 1.0 | 0.019 | 0.015 | 9.61 ( 4.4 / 5.2 ) | 1 | 0.391 ( 0.238 / 0.225 ) | 0.001 |
| 0.15 | 1.0 | 0.013 | 0.009 | 8.97 ( 4.3 / 4.7 ) | 1 | 0.357 ( 0.235 / 0.193 ) | 0.001 |
| 0.20 | 1.0 | 0.010 | 0.008 | 8.65 ( 4.2 / 4.5 ) | 1 | 0.340 ( 0.232 / 0.177 ) | 0.001 |
| 0.25 | 1.0 | 0.009 | 0.007 | 8.53 ( 4.2 / 4.4 ) | 1 | 0.334 ( 0.232 / 0.169 ) | 0.000 |
| 0.30 | 1.0 | 0.008 | 0.006 | 8.43 ( 4.2 / 4.3 ) | 1 | 0.328 ( 0.232 / 0.163 ) | 0.000 |
| 0.35 | 1.0 | 0.008 | 0.006 | 8.38 ( 4.2 / 4.2 ) | 1 | 0.324 ( 0.232 / 0.159 ) | 0.000 |
| 0.40 | 1.0 | 0.007 | 0.004 | 8.34 ( 4.2 / 4.2 ) | 1 | 0.321 ( 0.231 / 0.157 ) | 0.000 |
| 0.45 | 1.0 | 0.006 | 0.003 | 8.32 ( 4.2 / 4.2 ) | 1 | 0.318 ( 0.231 / 0.155 ) | 0.001 |
| 0.50 | 1.0 | 0.006 | 0.002 | 8.31 ( 4.2 / 4.2 ) | 1 | 0.317 ( 0.230 / 0.154 ) | 0.000 |

Table 4: Linear Regression: $n = 400$, $p = 400$

| Constant | Constant | Excess Risk | | $E[\widehat{J}]\left(E[\widehat{J}_1]/E[\widehat{J}_2]\right)$ | $P\{J_0 \subset \widehat{J}\}$ | $E\,\lvert\widehat{\alpha} - \alpha_0\rvert_1$ (on $J/J^c$) | $E\,\lvert\widehat{\tau} - \tau_0\rvert_1$ |
|---|---|---|---|---|---|---|---|
| for $\lambda_1$ | for $\lambda_2$ | Mean | Median | | | | |
| LASSO | | | | | | | |
| 0.05 | NA | 0.286 | 0.274 | 354.19 ( 209.3 / 144.9 ) | 1 | 18.802 ( 0.472 / 18.596 ) | 0.019 |
| 0.10 | NA | 0.249 | 0.243 | 294.99 ( 175.9 / 119.1 ) | 1 | 13.666 ( 0.432 / 13.469 ) | 0.012 |
| 0.15 | NA | 0.231 | 0.226 | 249.97 ( 149.7 / 100.2 ) | 1 | 10.373 ( 0.405 / 10.189 ) | 0.008 |
| 0.20 | NA | 0.225 | 0.222 | 213.11 ( 127.5 / 85.6 ) | 1 | 8.115 ( 0.391 / 7.940 ) | 0.005 |
| 0.25 | NA | 0.224 | 0.223 | 182.18 ( 108.4 / 73.8 ) | 1 | 6.488 ( 0.385 / 6.320 ) | 0.003 |
| 0.30 | NA | 0.227 | 0.227 | 156.07 ( 92.2 / 63.9 ) | 1 | 5.269 ( 0.385 / 5.105 ) | 0.002 |
| 0.35 | NA | 0.232 | 0.233 | 133.18 ( 78.0 / 55.1 ) | 1 | 4.326 ( 0.389 / 4.165 ) | 0.002 |
| 0.40 | NA | 0.239 | 0.241 | 113.50 ( 65.9 / 47.6 ) | 1 | 3.573 ( 0.393 / 3.416 ) | 0.001 |
| 0.45 | NA | 0.247 | 0.246 | 96.43 ( 55.3 / 41.1 ) | 1 | 2.972 ( 0.399 / 2.817 ) | 0.000 |
| 0.50 | NA | 0.257 | 0.255 | 81.42 ( 46.0 / 35.4 ) | 1 | 2.484 ( 0.407 / 2.331 ) | 0.000 |
| SCAD | | | | | | | |
| 0.05 | 0.5 | 0.052 | 0.047 | 84.32 ( 45.9 / 38.4 ) | 1 | 3.202 ( 0.288 / 3.007 ) | 0.019 |
| 0.10 | 0.5 | 0.038 | 0.035 | 81.27 ( 44.9 / 36.4 ) | 1 | 2.879 ( 0.275 / 2.691 ) | 0.012 |
| 0.15 | 0.5 | 0.030 | 0.031 | 79.12 ( 44.4 / 34.8 ) | 1 | 2.663 ( 0.263 / 2.482 ) | 0.008 |
| 0.20 | 0.5 | 0.024 | 0.023 | 78.30 ( 44.3 / 34.0 ) | 1 | 2.511 ( 0.255 / 2.336 ) | 0.005 |
| 0.25 | 0.5 | 0.019 | 0.017 | 77.83 ( 44.3 / 33.5 ) | 1 | 2.410 ( 0.251 / 2.238 ) | 0.003 |
| 0.30 | 0.5 | 0.017 | 0.016 | 77.82 ( 44.4 / 33.4 ) | 1 | 2.344 ( 0.249 / 2.173 ) | 0.002 |
| 0.35 | 0.5 | 0.015 | 0.013 | 77.81 ( 44.4 / 33.4 ) | 1 | 2.299 ( 0.248 / 2.128 ) | 0.002 |
| 0.40 | 0.5 | 0.012 | 0.011 | 77.86 ( 44.5 / 33.4 ) | 1 | 2.261 ( 0.246 / 2.091 ) | 0.001 |
| 0.45 | 0.5 | 0.011 | 0.008 | 77.91 ( 44.5 / 33.4 ) | 1 | 2.237 ( 0.245 / 2.068 ) | 0.000 |
| 0.50 | 0.5 | 0.010 | 0.009 | 77.95 ( 44.5 / 33.4 ) | 1 | 2.221 ( 0.244 / 2.052 ) | 0.000 |
| 0.05 | 1.0 | 0.033 | 0.028 | 11.70 ( 5.5 / 6.2 ) | 1 | 0.459 ( 0.269 / 0.276 ) | 0.019 |
| 0.10 | 1.0 | 0.022 | 0.018 | 10.96 ( 5.3 / 5.7 ) | 1 | 0.420 ( 0.259 / 0.241 ) | 0.012 |
| 0.15 | 1.0 | 0.016 | 0.016 | 10.47 ( 5.2 / 5.3 ) | 1 | 0.393 ( 0.249 / 0.220 ) | 0.008 |
| 0.20 | 1.0 | 0.011 | 0.009 | 10.16 ( 5.1 / 5.1 ) | 1 | 0.374 ( 0.243 / 0.207 ) | 0.005 |
| 0.25 | 1.0 | 0.008 | 0.004 | 10.01 ( 5.0 / 5.0 ) | 1 | 0.363 ( 0.240 / 0.198 ) | 0.003 |
| 0.30 | 1.0 | 0.007 | 0.004 | 9.91 ( 5.0 / 4.9 ) | 1 | 0.358 ( 0.238 / 0.193 ) | 0.002 |
| 0.35 | 1.0 | 0.006 | 0.001 | 9.86 ( 5.0 / 4.8 ) | 1 | 0.354 ( 0.238 / 0.190 ) | 0.002 |
| 0.40 | 1.0 | 0.004 | 0.001 | 9.83 ( 5.0 / 4.8 ) | 1 | 0.350 ( 0.236 / 0.187 ) | 0.001 |
| 0.45 | 1.0 | 0.004 | 0.000 | 9.82 ( 5.1 / 4.8 ) | 1 | 0.349 ( 0.236 / 0.186 ) | 0.000 |
| 0.50 | 1.0 | 0.004 | 0.000 | 9.80 ( 5.0 / 4.8 ) | 1 | 0.347 ( 0.235 / 0.184 ) | 0.000 |

Table 5: Logistic Regression: $n = 400$, $p = 50$

| Constant | Constant | Excess Risk | | $E[\widehat{J}]\left(E[\widehat{J}_1]/E[\widehat{J}_2]\right)$ | $P\{J_0 \subset \widehat{J}\}(\beta_1/\beta_3/\delta_2/\delta_3)$ | $E\left\|\widehat{\alpha}-\alpha_0\right\|_1$ (on $J/J^c$) | $E\left\|\widehat{\tau}-\tau_0\right\|_1$ |
|---|---|---|---|---|---|---|---|
| for $\lambda_1$ | for $\lambda_2$ | Mean | Median | | | | |
| LASSO | | | | | | | |
| 0.05 | NA | 0.123 | 0.121 | 69.09 ( 35.6 / 33.5 ) | 0.99 ( 1 / 1 / 1 / 0.99 ) | 12.913 ( 1.167 / 11.714 ) | 0.048 |
| 0.10 | NA | 0.057 | 0.055 | 48.13 ( 25.5 / 22.7 ) | 0.96 ( 1 / 1 / 0.99 / 0.97 ) | 6.217 ( 0.919 / 5.291 ) | 0.044 |
| 0.15 | NA | 0.036 | 0.034 | 33.59 ( 18.1 / 15.5 ) | 0.95 ( 1 / 1 / 0.98 / 0.96 ) | 3.818 ( 1.066 / 2.756 ) | 0.027 |
| 0.20 | NA | 0.028 | 0.026 | 23.45 ( 12.8 / 10.7 ) | 0.95 ( 1 / 1 / 0.99 / 0.96 ) | 2.758 ( 1.259 / 1.509 ) | 0.005 |
| 0.25 | NA | 0.026 | 0.025 | 16.62 ( 9.1 / 7.5 ) | 0.95 ( 1 / 1 / 0.99 / 0.96 ) | 2.291 ( 1.454 / 0.854 ) | 0.007 |
| 0.30 | NA | 0.027 | 0.026 | 11.80 ( 6.4 / 5.4 ) | 0.93 ( 1 / 1 / 0.99 / 0.95 ) | 2.094 ( 1.633 / 0.484 ) | 0.015 |
| 0.35 | NA | 0.030 | 0.030 | 8.57 ( 4.7 / 3.9 ) | 0.92 ( 1 / 1 / 0.98 / 0.93 ) | 2.050 ( 1.799 / 0.279 ) | 0.020 |
| 0.40 | NA | 0.035 | 0.034 | 6.56 ( 3.5 / 3.0 ) | 0.89 ( 1 / 1 / 0.98 / 0.92 ) | 2.086 ( 1.954 / 0.166 ) | 0.021 |
| 0.45 | NA | 0.039 | 0.039 | 5.32 ( 2.8 / 2.5 ) | 0.87 ( 1 / 1 / 0.97 / 0.9 ) | 2.168 ( 2.100 / 0.107 ) | 0.024 |
| 0.50 | NA | 0.045 | 0.044 | 4.56 ( 2.5 / 2.1 ) | 0.82 ( 1 / 1 / 0.96 / 0.86 ) | 2.276 ( 2.240 / 0.080 ) | 0.028 |
| SCAD | | | | | | | |
| 0.05 | 0.5 | 0.081 | 0.066 | 18.02 ( 7.3 / 10.7 ) | 0.97 ( 1 / 1 / 0.98 / 0.98 ) | 6.324 ( 1.529 / 4.804 ) | 0.049 |
| 0.10 | 0.5 | 0.048 | 0.044 | 11.28 ( 4.9 / 6.4 ) | 0.94 ( 1 / 1 / 0.97 / 0.96 ) | 3.519 ( 1.125 / 2.405 ) | 0.044 |
| 0.15 | 0.5 | 0.033 | 0.029 | 8.21 ( 3.8 / 4.4 ) | 0.91 ( 1 / 1 / 0.96 / 0.94 ) | 2.348 ( 1.018 / 1.341 ) | 0.027 |
| 0.20 | 0.5 | 0.024 | 0.020 | 6.58 ( 3.1 / 3.4 ) | 0.90 ( 1 / 1 / 0.97 / 0.92 ) | 1.681 ( 0.953 / 0.738 ) | 0.005 |
| 0.25 | 0.5 | 0.019 | 0.016 | 5.60 ( 2.8 / 2.8 ) | 0.89 ( 1 / 1 / 0.97 / 0.9 ) | 1.359 ( 0.939 / 0.431 ) | 0.007 |
| 0.30 | 0.5 | 0.016 | 0.013 | 4.99 ( 2.5 / 2.5 ) | 0.86 ( 1 / 1 / 0.97 / 0.88 ) | 1.204 ( 0.959 / 0.260 ) | 0.015 |
| 0.35 | 0.5 | 0.015 | 0.012 | 4.60 ( 2.4 / 2.2 ) | 0.83 ( 1 / 1 / 0.96 / 0.86 ) | 1.114 ( 0.970 / 0.158 ) | 0.020 |
| 0.40 | 0.5 | 0.014 | 0.011 | 4.37 ( 2.3 / 2.1 ) | 0.80 ( 1 / 1 / 0.96 / 0.84 ) | 1.088 ( 1.004 / 0.098 ) | 0.021 |
| 0.45 | 0.5 | 0.015 | 0.012 | 4.22 ( 2.3 / 1.9 ) | 0.76 ( 1 / 1 / 0.95 / 0.8 ) | 1.109 ( 1.054 / 0.069 ) | 0.024 |
| 0.50 | 0.5 | 0.016 | 0.014 | 4.13 ( 2.3 / 1.9 ) | 0.73 ( 1 / 1 / 0.95 / 0.77 ) | 1.164 ( 1.126 / 0.054 ) | 0.028 |
| | | | | | | | |
| 0.05 | 1.0 | 0.039 | 0.033 | 8.89 ( 2.9 / 6.0 ) | 0.94 ( 1 / 1 / 0.97 / 0.97 ) | 2.935 ( 1.064 / 1.875 ) | 0.048 |
| 0.10 | 1.0 | 0.022 | 0.018 | 5.45 ( 2.3 / 3.1 ) | 0.87 ( 1 / 1 / 0.95 / 0.91 ) | 1.590 ( 0.989 / 0.612 ) | 0.044 |
| 0.15 | 1.0 | 0.017 | 0.013 | 4.42 ( 2.1 / 2.3 ) | 0.82 ( 1 / 1 / 0.93 / 0.87 ) | 1.234 ( 1.007 / 0.242 ) | 0.027 |
| 0.20 | 1.0 | 0.015 | 0.012 | 3.99 ( 2.1 / 1.9 ) | 0.77 ( 1 / 1 / 0.93 / 0.81 ) | 1.118 ( 1.045 / 0.087 ) | 0.005 |
| 0.25 | 1.0 | 0.016 | 0.013 | 3.82 ( 2.0 / 1.8 ) | 0.73 ( 1 / 1 / 0.93 / 0.77 ) | 1.148 ( 1.114 / 0.049 ) | 0.007 |
| 0.30 | 1.0 | 0.017 | 0.014 | 3.70 ( 2.0 / 1.7 ) | 0.66 ( 1 / 0.99 / 0.91 / 0.73 ) | 1.227 ( 1.208 / 0.037 ) | 0.015 |
| 0.35 | 1.0 | 0.018 | 0.016 | 3.60 ( 2.0 / 1.6 ) | 0.60 ( 1 / 0.99 / 0.9 / 0.68 ) | 1.330 ( 1.319 / 0.030 ) | 0.020 |
| 0.40 | 1.0 | 0.021 | 0.018 | 3.49 ( 2.0 / 1.5 ) | 0.50 ( 1 / 0.99 / 0.87 / 0.61 ) | 1.457 ( 1.450 / 0.028 ) | 0.021 |
| 0.45 | 1.0 | 0.023 | 0.021 | 3.38 ( 2.0 / 1.4 ) | 0.42 ( 1 / 0.99 / 0.83 / 0.55 ) | 1.583 ( 1.577 / 0.029 ) | 0.024 |
| 0.50 | 1.0 | 0.026 | 0.025 | 3.26 ( 2.0 / 1.3 ) | 0.34 ( 1 / 0.98 / 0.78 / 0.49 ) | 1.722 ( 1.717 / 0.031 ) | 0.028 |

Table 6: Logistic Regression: $n = 400$, $p = 100$

| Constant for $\lambda_1$ | Constant for $\lambda_2$ | Excess Risk Mean | Excess Risk Median | $E[\widehat{J}]$ $\left(E[\widehat{J}_1]/E[\widehat{J}_2]\right)$ | $P\{J_0 \subset \widehat{J}\}(\beta_1/\beta_3/\delta_2/\delta_3)$ | $E\left\|\widehat{\alpha} - \alpha_0\right\|_1$ (on $J/J^c$) | $E\left\|\widehat{\tau} - \tau_0\right\|_1$ |
|---|---|---|---|---|---|---|---|
| LASSO | | | | | | | |
| 0.05 | NA | 0.230 | 0.227 | 115.73 ( 62.8 / 53.0 ) | 0.99 ( 1 / 1 / 1 / 1 ) | 21.933 ( 1.423 / 20.620 ) | 0.002 |
| 0.10 | NA | 0.088 | 0.086 | 78.13 ( 43.8 / 34.3 ) | 0.97 ( 1 / 1 / 0.99 / 0.98 ) | 9.372 ( 0.952 / 8.459 ) | 0.021 |
| 0.15 | NA | 0.049 | 0.048 | 52.97 ( 30.2 / 22.8 ) | 0.94 ( 1 / 1 / 0.98 / 0.96 ) | 5.376 ( 1.132 / 4.268 ) | 0.011 |
| 0.20 | NA | 0.036 | 0.034 | 35.77 ( 20.4 / 15.4 ) | 0.92 ( 1 / 1 / 0.97 / 0.94 ) | 3.671 ( 1.367 / 2.325 ) | 0.001 |
| 0.25 | NA | 0.031 | 0.030 | 24.02 ( 13.6 / 10.4 ) | 0.90 ( 1 / 1 / 0.97 / 0.92 ) | 2.837 ( 1.572 / 1.287 ) | 0.011 |
| 0.30 | NA | 0.031 | 0.030 | 15.99 ( 9.0 / 7.0 ) | 0.90 ( 1 / 1 / 0.98 / 0.92 ) | 2.428 ( 1.745 / 0.709 ) | 0.015 |
| 0.35 | NA | 0.034 | 0.033 | 10.77 ( 6.0 / 4.8 ) | 0.88 ( 1 / 1 / 0.97 / 0.9 ) | 2.268 ( 1.909 / 0.390 ) | 0.022 |
| 0.40 | NA | 0.038 | 0.037 | 7.67 ( 4.2 / 3.5 ) | 0.86 ( 1 / 1 / 0.96 / 0.88 ) | 2.249 ( 2.063 / 0.222 ) | 0.023 |
| 0.45 | NA | 0.043 | 0.043 | 5.77 ( 3.1 / 2.6 ) | 0.83 ( 1 / 1 / 0.95 / 0.86 ) | 2.299 ( 2.207 / 0.134 ) | 0.023 |
| 0.50 | NA | 0.049 | 0.049 | 4.73 ( 2.6 / 2.2 ) | 0.80 ( 1 / 1 / 0.94 / 0.85 ) | 2.390 ( 2.345 / 0.094 ) | 0.028 |
| SCAD | | | | | | | |
| 0.05 | 0.5 | 0.132 | 0.124 | 27.14 ( 11.8 / 15.4 ) | 0.98 ( 1 / 1 / 0.99 / 0.99 ) | 9.026 ( 1.537 / 7.571 ) | 0.002 |
| 0.10 | 0.5 | 0.069 | 0.064 | 14.62 ( 6.5 / 8.1 ) | 0.93 ( 1 / 1 / 0.97 / 0.96 ) | 4.637 ( 1.169 / 3.509 ) | 0.021 |
| 0.15 | 0.5 | 0.043 | 0.039 | 9.75 ( 4.5 / 5.2 ) | 0.89 ( 1 / 1 / 0.96 / 0.91 ) | 2.814 ( 1.048 / 1.791 ) | 0.011 |
| 0.20 | 0.5 | 0.029 | 0.025 | 7.48 ( 3.6 / 3.9 ) | 0.86 ( 1 / 1 / 0.96 / 0.88 ) | 1.964 ( 1.002 / 0.980 ) | 0.001 |
| 0.25 | 0.5 | 0.022 | 0.019 | 6.09 ( 3.0 / 3.0 ) | 0.83 ( 1 / 1 / 0.95 / 0.86 ) | 1.535 ( 0.996 / 0.556 ) | 0.011 |
| 0.30 | 0.5 | 0.018 | 0.015 | 5.28 ( 2.7 / 2.6 ) | 0.80 ( 1 / 1 / 0.95 / 0.83 ) | 1.305 ( 1.004 / 0.316 ) | 0.015 |
| 0.35 | 0.5 | 0.017 | 0.013 | 4.77 ( 2.5 / 2.3 ) | 0.77 ( 1 / 1 / 0.95 / 0.81 ) | 1.210 ( 1.030 / 0.194 ) | 0.022 |
| 0.40 | 0.5 | 0.016 | 0.013 | 4.44 ( 2.4 / 2.1 ) | 0.74 ( 1 / 1 / 0.94 / 0.78 ) | 1.193 ( 1.075 / 0.131 ) | 0.023 |
| 0.45 | 0.5 | 0.016 | 0.013 | 4.29 ( 2.3 / 1.9 ) | 0.71 ( 1 / 1 / 0.93 / 0.75 ) | 1.221 ( 1.143 / 0.092 ) | 0.023 |
| 0.50 | 0.5 | 0.017 | 0.014 | 4.20 ( 2.3 / 1.9 ) | 0.65 ( 1 / 1 / 0.93 / 0.71 ) | 1.292 ( 1.237 / 0.071 ) | 0.028 |
| 0.05 | 1.0 | 0.050 | 0.045 | 11.42 ( 3.8 / 7.7 ) | 0.95 ( 1 / 1 / 0.98 / 0.98 ) | 3.487 ( 1.007 / 2.509 ) | 0.002 |
| 0.10 | 1.0 | 0.024 | 0.020 | 5.67 ( 2.3 / 3.3 ) | 0.85 ( 1 / 1 / 0.94 / 0.9 ) | 1.635 ( 0.977 / 0.673 ) | 0.021 |
| 0.15 | 1.0 | 0.019 | 0.015 | 4.38 ( 2.1 / 2.2 ) | 0.76 ( 1 / 1 / 0.92 / 0.82 ) | 1.280 ( 1.058 / 0.236 ) | 0.011 |
| 0.20 | 1.0 | 0.018 | 0.014 | 3.92 ( 2.0 / 1.9 ) | 0.69 ( 1 / 1 / 0.9 / 0.76 ) | 1.248 ( 1.153 / 0.111 ) | 0.001 |
| 0.25 | 1.0 | 0.019 | 0.015 | 3.70 ( 2.0 / 1.7 ) | 0.62 ( 1 / 0.99 / 0.89 / 0.69 ) | 1.304 ( 1.263 / 0.059 ) | 0.011 |
| 0.30 | 1.0 | 0.020 | 0.016 | 3.55 ( 2.0 / 1.5 ) | 0.55 ( 1 / 0.99 / 0.86 / 0.63 ) | 1.400 ( 1.377 / 0.041 ) | 0.015 |
| 0.35 | 1.0 | 0.022 | 0.019 | 3.43 ( 2.0 / 1.4 ) | 0.46 ( 1 / 0.99 / 0.84 / 0.56 ) | 1.525 ( 1.511 / 0.035 ) | 0.022 |
| 0.40 | 1.0 | 0.024 | 0.022 | 3.30 ( 2.0 / 1.3 ) | 0.37 ( 1 / 0.98 / 0.79 / 0.49 ) | 1.666 ( 1.654 / 0.036 ) | 0.023 |
| 0.45 | 1.0 | 0.027 | 0.025 | 3.18 ( 2.0 / 1.2 ) | 0.29 ( 1 / 0.98 / 0.72 / 0.45 ) | 1.800 ( 1.791 / 0.037 ) | 0.023 |
| 0.50 | 1.0 | 0.031 | 0.029 | 3.03 ( 2.0 / 1.0 ) | 0.22 ( 1 / 0.98 / 0.66 / 0.39 ) | 1.961 ( 1.955 / 0.036 ) | 0.028 |

Table 7: Logistic Regression: $n = 400$, $p = 200$

| Constant | Constant | Excess Risk | | $E[\widehat{J}]$ $\left(E[\widehat{J}_1]/E[\widehat{J}_2]\right)$ | $P\{J_0 \subset \widehat{J}\}(\beta_1/\beta_3/\delta_2/\delta_3)$ | $E\left\|\widehat{\alpha}-\alpha_0\right\|_1$ (on $J/J^c$) | $E\left\|\widehat{\tau}-\tau_0\right\|_1$ |
|---|---|---|---|---|---|---|---|
| for $\lambda_1$ | for $\lambda_2$ | Mean | Median | | | | |
| <u>LASSO</u> | | | | | | | |
| 0.05 | NA | 0.318 | 0.315 | 157.22 ( 91.3 / 65.9 ) | 0.98 ( 1 / 1 / 0.99 / 0.99 ) | 26.697 ( 1.499 / 25.441 ) | 0.074 |
| 0.10 | NA | 0.123 | 0.122 | 112.42 ( 66.6 / 45.8 ) | 0.97 ( 1 / 1 / 0.99 / 0.98 ) | 12.250 ( 1.021 / 11.319 ) | 0.047 |
| 0.15 | NA | 0.065 | 0.064 | 77.20 ( 45.7 / 31.5 ) | 0.95 ( 1 / 1 / 0.98 / 0.96 ) | 7.012 ( 1.208 / 5.854 ) | 0.037 |
| 0.20 | NA | 0.044 | 0.043 | 51.88 ( 30.6 / 21.3 ) | 0.92 ( 1 / 1 / 0.98 / 0.93 ) | 4.626 ( 1.456 / 3.206 ) | 0.031 |
| 0.25 | NA | 0.036 | 0.035 | 33.59 ( 19.7 / 13.9 ) | 0.89 ( 1 / 1 / 0.97 / 0.9 ) | 3.395 ( 1.664 / 1.763 ) | 0.025 |
| 0.30 | NA | 0.035 | 0.034 | 21.32 ( 12.3 / 9.0 ) | 0.87 ( 1 / 1 / 0.97 / 0.89 ) | 2.756 ( 1.838 / 0.950 ) | 0.025 |
| 0.35 | NA | 0.037 | 0.036 | 13.38 ( 7.6 / 5.8 ) | 0.85 ( 1 / 1 / 0.96 / 0.88 ) | 2.457 ( 1.994 / 0.499 ) | 0.027 |
| 0.40 | NA | 0.041 | 0.041 | 8.73 ( 4.9 / 3.8 ) | 0.83 ( 1 / 1 / 0.95 / 0.86 ) | 2.368 ( 2.144 / 0.266 ) | 0.026 |
| 0.45 | NA | 0.047 | 0.046 | 6.15 ( 3.4 / 2.7 ) | 0.79 ( 1 / 1 / 0.93 / 0.84 ) | 2.396 ( 2.290 / 0.153 ) | 0.027 |
| 0.50 | NA | 0.053 | 0.052 | 4.81 ( 2.6 / 2.2 ) | 0.74 ( 1 / 1 / 0.9 / 0.82 ) | 2.480 ( 2.430 / 0.103 ) | 0.030 |
| <u>SCAD</u> | | | | | | | |
| 0.05 | 0.5 | 0.204 | 0.197 | 32.44 ( 16.1 / 16.4 ) | 0.94 ( 1 / 1 / 0.96 / 0.97 ) | 11.373 ( 1.620 / 9.923 ) | 0.074 |
| 0.10 | 0.5 | 0.090 | 0.085 | 17.53 ( 8.5 / 9.0 ) | 0.92 ( 1 / 1 / 0.96 / 0.95 ) | 5.198 ( 1.158 / 4.115 ) | 0.047 |
| 0.15 | 0.5 | 0.049 | 0.046 | 11.26 ( 5.5 / 5.7 ) | 0.87 ( 1 / 1 / 0.96 / 0.9 ) | 2.980 ( 1.056 / 1.966 ) | 0.037 |
| 0.20 | 0.5 | 0.033 | 0.030 | 8.22 ( 4.1 / 4.1 ) | 0.81 ( 1 / 1 / 0.95 / 0.85 ) | 2.094 ( 1.045 / 1.078 ) | 0.031 |
| 0.25 | 0.5 | 0.025 | 0.021 | 6.55 ( 3.3 / 3.2 ) | 0.78 ( 1 / 1 / 0.94 / 0.82 ) | 1.693 ( 1.074 / 0.642 ) | 0.025 |
| 0.30 | 0.5 | 0.021 | 0.017 | 5.55 ( 2.9 / 2.7 ) | 0.76 ( 1 / 1 / 0.94 / 0.8 ) | 1.443 ( 1.085 / 0.377 ) | 0.025 |
| 0.35 | 0.5 | 0.019 | 0.016 | 4.89 ( 2.6 / 2.3 ) | 0.72 ( 1 / 1 / 0.93 / 0.77 ) | 1.351 ( 1.135 / 0.234 ) | 0.027 |
| 0.40 | 0.5 | 0.018 | 0.015 | 4.52 ( 2.4 / 2.1 ) | 0.69 ( 1 / 1 / 0.91 / 0.75 ) | 1.330 ( 1.187 / 0.160 ) | 0.026 |
| 0.45 | 0.5 | 0.019 | 0.016 | 4.31 ( 2.4 / 1.9 ) | 0.64 ( 1 / 1 / 0.91 / 0.71 ) | 1.357 ( 1.260 / 0.116 ) | 0.027 |
| 0.50 | 0.5 | 0.020 | 0.017 | 4.21 ( 2.4 / 1.8 ) | 0.59 ( 1 / 0.99 / 0.9 / 0.67 ) | 1.418 ( 1.349 / 0.089 ) | 0.030 |
| 0.05 | 1.0 | 0.056 | 0.053 | 11.26 ( 4.5 / 6.8 ) | 0.86 ( 1 / 1 / 0.93 / 0.93 ) | 3.259 ( 1.029 / 2.277 ) | 0.074 |
| 0.10 | 1.0 | 0.023 | 0.020 | 5.31 ( 2.4 / 2.9 ) | 0.79 ( 1 / 1 / 0.91 / 0.85 ) | 1.487 ( 1.049 / 0.459 ) | 0.047 |
| 0.15 | 1.0 | 0.020 | 0.016 | 4.17 ( 2.1 / 2.1 ) | 0.70 ( 1 / 1 / 0.9 / 0.78 ) | 1.290 ( 1.152 / 0.157 ) | 0.037 |
| 0.20 | 1.0 | 0.020 | 0.018 | 3.76 ( 2.0 / 1.7 ) | 0.58 ( 1 / 0.99 / 0.86 / 0.68 ) | 1.358 ( 1.295 / 0.083 ) | 0.031 |
| 0.25 | 1.0 | 0.022 | 0.019 | 3.56 ( 2.0 / 1.5 ) | 0.51 ( 0.99 / 0.99 / 0.82 / 0.62 ) | 1.487 ( 1.444 / 0.064 ) | 0.025 |
| 0.30 | 1.0 | 0.024 | 0.021 | 3.40 ( 2.0 / 1.4 ) | 0.44 ( 1 / 0.99 / 0.78 / 0.55 ) | 1.601 ( 1.581 / 0.043 ) | 0.025 |
| 0.35 | 1.0 | 0.026 | 0.023 | 3.27 ( 2.0 / 1.3 ) | 0.35 ( 1 / 0.98 / 0.76 / 0.49 ) | 1.713 ( 1.700 / 0.039 ) | 0.027 |
| 0.40 | 1.0 | 0.028 | 0.026 | 3.16 ( 2.0 / 1.2 ) | 0.28 ( 1 / 0.98 / 0.72 / 0.43 ) | 1.848 ( 1.836 / 0.041 ) | 0.026 |
| 0.45 | 1.0 | 0.032 | 0.030 | 3.04 ( 2.0 / 1.1 ) | 0.22 ( 1 / 0.97 / 0.67 / 0.38 ) | 1.996 ( 1.986 / 0.042 ) | 0.027 |
| 0.50 | 1.0 | 0.037 | 0.035 | 2.88 ( 2.0 / 0.9 ) | 0.15 ( 0.99 / 0.96 / 0.59 / 0.32 ) | 2.160 ( 2.154 / 0.042 ) | 0.030 |

Table 8: Logistic Regression: $n = 400$, $p = 400$

| Constant for $\lambda_1$ | Constant for $\lambda_2$ | Excess Risk | | $E[\widehat{J}]\ \left(E[\widehat{J}_1]/E[\widehat{J}_2]\right)$ | $P\{J_0 \subset \widehat{J}\}(\beta_1/\beta_3/\delta_2/\delta_3)$ | $E\left\|\widehat{\alpha}-\alpha_0\right\|_1$ (on $J/J^c$) | $E\left\|\widehat{\tau}-\tau_0\right\|_1$ |
|---|---|---|---|---|---|---|---|
| | | Mean | Median | | | | |
| LASSO | | | | | | | |
| 0.05 | NA | 0.307 | 0.306 | 182.99 ( 112.0 / 71.0 ) | 0.86 ( 1 / 1 / 0.93 / 0.93 ) | 25.418 ( 1.590 / 24.107 ) | 0.144 |
| 0.10 | NA | 0.145 | 0.143 | 142.37 ( 86.7 / 55.6 ) | 0.88 ( 1 / 1 / 0.95 / 0.93 ) | 14.069 ( 1.296 / 12.902 ) | 0.127 |
| 0.15 | NA | 0.081 | 0.081 | 103.20 ( 62.7 / 40.5 ) | 0.87 ( 1 / 1 / 0.95 / 0.91 ) | 8.622 ( 1.433 / 7.262 ) | 0.104 |
| 0.20 | NA | 0.054 | 0.053 | 70.83 ( 42.9 / 27.9 ) | 0.87 ( 1 / 1 / 0.95 / 0.91 ) | 5.681 ( 1.617 / 4.114 ) | 0.078 |
| 0.25 | NA | 0.043 | 0.042 | 45.87 ( 27.5 / 18.3 ) | 0.86 ( 1 / 1 / 0.95 / 0.9 ) | 4.022 ( 1.779 / 2.283 ) | 0.055 |
| 0.30 | NA | 0.040 | 0.039 | 28.16 ( 16.6 / 11.6 ) | 0.85 ( 1 / 1 / 0.95 / 0.89 ) | 3.108 ( 1.929 / 1.218 ) | 0.042 |
| 0.35 | NA | 0.041 | 0.041 | 16.89 ( 9.8 / 7.1 ) | 0.84 ( 1 / 1 / 0.95 / 0.88 ) | 2.657 ( 2.072 / 0.625 ) | 0.033 |
| 0.40 | NA | 0.045 | 0.045 | 10.18 ( 5.8 / 4.4 ) | 0.81 ( 1 / 1 / 0.94 / 0.85 ) | 2.486 ( 2.215 / 0.316 ) | 0.026 |
| 0.45 | NA | 0.051 | 0.050 | 6.71 ( 3.8 / 3.0 ) | 0.78 ( 1 / 1 / 0.91 / 0.84 ) | 2.475 ( 2.357 / 0.169 ) | 0.028 |
| 0.50 | NA | 0.057 | 0.057 | 4.96 ( 2.8 / 2.2 ) | 0.71 ( 1 / 1 / 0.88 / 0.8 ) | 2.547 ( 2.497 / 0.107 ) | 0.030 |
| SCAD | | | | | | | |
| 0.05 | 0.5 | 0.213 | 0.205 | 30.12 ( 16.2 / 13.9 ) | 0.77 ( 1 / 1 / 0.88 / 0.87 ) | 10.439 ( 1.613 / 9.025 ) | 0.144 |
| 0.10 | 0.5 | 0.098 | 0.092 | 18.14 ( 9.4 / 8.8 ) | 0.75 ( 1 / 0.99 / 0.88 / 0.85 ) | 5.113 ( 1.269 / 3.934 ) | 0.127 |
| 0.15 | 0.5 | 0.054 | 0.049 | 11.99 ( 6.1 / 5.9 ) | 0.75 ( 1 / 1 / 0.89 / 0.83 ) | 3.068 ( 1.170 / 1.949 ) | 0.104 |
| 0.20 | 0.5 | 0.036 | 0.033 | 8.65 ( 4.4 / 4.3 ) | 0.73 ( 1 / 1 / 0.9 / 0.8 ) | 2.199 ( 1.156 / 1.077 ) | 0.078 |
| 0.25 | 0.5 | 0.027 | 0.024 | 6.79 ( 3.5 / 3.3 ) | 0.71 ( 1 / 1 / 0.9 / 0.78 ) | 1.745 ( 1.150 / 0.621 ) | 0.055 |
| 0.30 | 0.5 | 0.022 | 0.019 | 5.67 ( 3.0 / 2.7 ) | 0.70 ( 1 / 1 / 0.9 / 0.76 ) | 1.505 ( 1.162 / 0.364 ) | 0.042 |
| 0.35 | 0.5 | 0.020 | 0.017 | 4.98 ( 2.7 / 2.3 ) | 0.67 ( 1 / 1 / 0.91 / 0.73 ) | 1.408 ( 1.209 / 0.218 ) | 0.033 |
| 0.40 | 0.5 | 0.019 | 0.016 | 4.57 ( 2.5 / 2.1 ) | 0.64 ( 1 / 1 / 0.9 / 0.7 ) | 1.375 ( 1.255 / 0.137 ) | 0.026 |
| 0.45 | 0.5 | 0.020 | 0.017 | 4.35 ( 2.4 / 1.9 ) | 0.58 ( 1 / 1 / 0.89 / 0.67 ) | 1.406 ( 1.331 / 0.094 ) | 0.028 |
| 0.50 | 0.5 | 0.021 | 0.019 | 4.20 ( 2.4 / 1.8 ) | 0.52 ( 1 / 0.99 / 0.88 / 0.62 ) | 1.475 ( 1.423 / 0.073 ) | 0.030 |
| 0.05 | 1.0 | 0.047 | 0.044 | 8.08 ( 3.7 / 4.4 ) | 0.63 ( 1 / 0.99 / 0.8 / 0.8 ) | 2.474 ( 1.263 / 1.254 ) | 0.144 |
| 0.10 | 1.0 | 0.029 | 0.027 | 4.57 ( 2.3 / 2.2 ) | 0.53 ( 1 / 0.99 / 0.76 / 0.71 ) | 1.635 ( 1.362 / 0.301 ) | 0.127 |
| 0.15 | 1.0 | 0.026 | 0.024 | 3.72 ( 2.1 / 1.6 ) | 0.48 ( 1 / 0.99 / 0.74 / 0.64 ) | 1.533 ( 1.450 / 0.108 ) | 0.104 |
| 0.20 | 1.0 | 0.025 | 0.023 | 3.45 ( 2.0 / 1.4 ) | 0.43 ( 1 / 0.98 / 0.74 / 0.57 ) | 1.569 ( 1.535 / 0.060 ) | 0.078 |
| 0.25 | 1.0 | 0.026 | 0.024 | 3.33 ( 2.0 / 1.3 ) | 0.40 ( 1 / 0.98 / 0.73 / 0.55 ) | 1.630 ( 1.616 / 0.040 ) | 0.055 |
| 0.30 | 1.0 | 0.027 | 0.025 | 3.22 ( 2.0 / 1.2 ) | 0.34 ( 1 / 0.99 / 0.72 / 0.48 ) | 1.732 ( 1.727 / 0.032 ) | 0.042 |
| 0.35 | 1.0 | 0.029 | 0.027 | 3.12 ( 2.0 / 1.1 ) | 0.28 ( 1 / 0.98 / 0.68 / 0.44 ) | 1.844 ( 1.840 / 0.034 ) | 0.033 |
| 0.40 | 1.0 | 0.032 | 0.031 | 3.00 ( 2.0 / 1.0 ) | 0.20 ( 1 / 0.97 / 0.63 / 0.39 ) | 1.978 ( 1.976 / 0.035 ) | 0.026 |
| 0.45 | 1.0 | 0.036 | 0.035 | 2.87 ( 2.0 / 0.9 ) | 0.14 ( 1 / 0.97 / 0.57 / 0.33 ) | 2.138 ( 2.137 / 0.037 ) | 0.028 |
| 0.50 | 1.0 | 0.041 | 0.040 | 2.73 ( 2.0 / 0.8 ) | 0.09 ( 0.99 / 0.96 / 0.49 / 0.28 ) | 2.295 ( 2.294 / 0.043 ) | 0.030 |

Then, due to the Hölder inequality,

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

Furthermore, the Bernstein's inequality and the fact that $X_{ij}(\tau)$ is a step function with less than $n$ jump points yield that

$$\mathrm{E}\left( \sup_{\tau} \max_{j\leq 2p} \left| \frac{1}{n}\sum_{i=1}^{n} \eta_i X_{ij}(\tau) \right| \right) \leq c_{np}.$$

Thus,

$$\Pr\left\{ \sup_{\tau} \sup_{|\alpha-\alpha_0|_1\leq m_{1n}} |\nu_n(\alpha,\tau)-\nu_n(\alpha_0,\tau)| > a_n m_{1n} \right\} \leq (a_n m_{1n})^{-1} 2L m_{1n} c_{np} = \delta.$$

**Proof of** $(A.2)$ : Let $\{\eta_i\}$ be a iid Rademacher sequence, independent of $\{X_i\}$. By the symmetrization theorem, and for $k \geq \log_2(m_{2n}/m_{1n})$, and then by the contraction theorem,

$$\begin{aligned}
&\mathrm{E}\sup_{\tau} \sup_{m_{1n}\leq|\alpha-\alpha_0|_1\leq m_{2n}} \frac{|\nu_n(\alpha,\tau)-\nu_n(\alpha_0,\tau)|}{|\alpha-\alpha_0|_1} \\
&\leq 2\mathrm{E}\left( \sup_{\tau} \sup_{m_{1n}\leq|\alpha-\alpha_0|_1\leq m_{2n}} \left| \frac{1}{n}\sum_{i=1}^{n} \eta_i \frac{\rho\left(Y_i, X_i(\tau)^T\alpha\right) - \rho\left(Y_i, X_i(\tau)^T\alpha_0\right)}{|\alpha-\alpha_0|_1} \right| \right) \\
&\leq 2\sum_{j=1}^{k} \mathrm{E}\left( \sup_{\tau} \sup_{2^{j-1}m_{1n}\leq|\alpha-\alpha_0|_1\leq 2^j m_{1n}} \left| \frac{1}{n}\sum_{i=1}^{n} \eta_i \frac{\rho\left(Y_i, X_i(\tau)^T\alpha\right) - \rho\left(Y_i, X_i(\tau)^T\alpha_0\right)}{2^{j-1}m_{1n}} \right| \right) \\
&\leq 4L\sum_{j=1}^{k} \mathrm{E}\left( \sup_{\tau} \sup_{2^{j-1}m_{1n}\leq|\alpha-\alpha_0|_1\leq 2^j m_{1n}} \left| \frac{1}{n}\sum_{i=1}^{n} \eta_i \frac{X_i(\tau)^T(\alpha-\alpha_0)}{2^{j-1}m_{1n}} \right| \right).
\end{aligned}$$

Next, due to the Hölder inequality,

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

Then, by the Bernstein's inequality and the Markov inequality,

$$\Pr\left\{ \sup_{\tau} \sup_{m_{1n}\leq|\alpha-\alpha_0|_1\leq m_{2n}} \frac{|\nu_n(\alpha,\tau)-\nu_n(\alpha_0,\tau)|}{|\alpha-\alpha_0|_1} > b_n \right\} \leq b_n^{-1} 8Lkc_{np} = \delta.$$

**Proof of** $(A.3)$ : As above, by the symmetrization and contraction,

$$\begin{aligned}
& \mathrm{E}\sup_{\tau\in\mathcal{T}_\eta}\left|\nu_n\left(\alpha_0,\tau\right)-\nu_n\left(\alpha_0,\tau_0\right)\right| \\
\leq\ & 2\mathrm{E}\sup_{\tau\in\mathcal{T}_\eta}\left|\frac{1}{n}\sum_{i=1}^{n}\eta_i\rho\left(Y_i,X_i\left(\tau\right)^T\alpha_0\right)-\rho\left(Y_i,X_i\left(\tau_0\right)^T\alpha_0\right)\right| \\
\leq\ & 4L\mathrm{E}\sup_{\tau\in\mathcal{T}_\eta}\left|\frac{1}{n}\sum_{i=1}^{n}\eta_iX_i^T\delta_0\left(1\left\{Q_i>\tau\right\}-1\left\{Q_i>\tau_0\right\}\right)\right| \\
\leq\ & \frac{4LK_2\eta C_1}{\sqrt{n}},
\end{aligned}$$

where the last inequality is due to Theorem 2.14.1 of van der Vaart and Wellner (1996) with $K_2$ in Assumption 2.1 and $C_1$ given in their theorem.

## A.1 Proof of Theorem 2.1

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

$$\begin{aligned}
R(\widehat{\alpha},\widehat{\tau}) &\leq [\nu(\tau_0,\alpha_0)-\nu_n(\widehat{\tau},\widehat{\alpha})]+\lambda_n|D_0\alpha_0|_1-\lambda_n|\widehat{D}\widehat{\alpha}|_1 \\
&= [\nu(\widehat{\tau},\alpha_0)-\nu_n(\widehat{\tau},\widehat{\alpha})]+[\nu(\tau_0,\alpha_0)-\nu_n(\widehat{\tau},\alpha_0)] \qquad (A.4)\\
&\quad +\lambda_n\left(|D_0\alpha_0|_1-|\widehat{D}\widehat{\alpha}|_1\right).
\end{aligned}$$

Note that the second component $[\nu(\tau_0,\alpha_0)-\nu_n(\widehat{\tau},\alpha_0)]=o_p\left(n^{-1/2}\log n\right)$ due to Lemma A.1. Thus, we focus on the other two terms in the following discussion.

Suppose that $\left|\widehat{\alpha}-\alpha_0\right|_1\leq\left|\alpha_0\right|_1$. 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 Lemma A.1 with $m_{1n}=\left|\alpha_0\right|_1$ and $m_{2n}=2Mp$. Then, from $(A.1)$

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

w.p.1 due to Lemma A.1. Thus, the theorem follows.

Now assume that $\left|\widehat{\alpha}-\alpha_0\right|_1>\left|\alpha_0\right|_1$. Then, due to Lemma A.1, w.p.1, (since $b_n=D_{\min}\lambda_n$ and $\underline{D}\left|\widehat{\alpha}-\alpha_0\right|_1\leq\left|\widehat{D}\left(\widehat{\alpha}-\alpha_0\right)\right|_1$)

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

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

## A.2 Proof of Theorem 2.2

Note 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 $\tau > \tau_0$,

$$R(\alpha, \tau) = E\left(\rho\left(Y, X^T\beta\right) - \rho\left(Y, X^T\beta_0\right)\right) 1\left\{Q \le \tau_0\right\} + E\left(\rho\left(Y, X^T\theta\right) - \rho\left(Y, X^T\theta_0\right)\right) 1\left\{Q > \tau\right\} \tag{A.5}$$
$$+ E\left(\rho\left(Y, X^T\beta\right) - \rho\left(Y, X^T\theta_0\right)\right) 1\left\{\tau_0 < Q \le \tau\right\}.$$

Then identification condition $E[\rho(Y, X(\tau_0)^T\alpha) - \rho(Y, X(\tau_0)^T\alpha_0)|Q] \ge 0$ implies that all the three terms on the right hand side (RHS) are nonnegative. To see this, the first two terms are nonnegative by simply multiplying $E[\rho(Y, X(\tau_0)^T\alpha) - \rho(Y, X(\tau_0)^T\alpha_0)|Q] \ge 0$ with $1\{Q \le \tau_0\}$ 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 $1\{\tau_0 < Q \le \tau\}E[\rho(Y, X(\tau_0)^T\alpha) - \rho(Y, X(\tau_0)^T\alpha_0)|Q] \ge 0$. We then have,

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

which yields the non-negativeness of the third term. Thus all the three terms on the RHS of (A.5) are bounded by $R(\alpha, \tau)$.

We now prove that, for any $\epsilon' > 0$, there is $\varepsilon > 0$ and $C_\varepsilon > 0$ so that for all $\tau > \tau_0$, and $\alpha \in \mathbb{R}^{2p}$, the fact that $R(\alpha, \tau) < \min\{C_\varepsilon/2, \varepsilon\}$ implies that $\tau < \tau_0 + \epsilon'$.

Applying the triangle inequality, for all $\beta$ and $\tau > \tau_0$,

$$E\left(\rho\left(Y, X^T\beta_0\right) - \rho\left(Y, X^T\theta_0\right)\right) 1\left\{\tau_0 < Q \le \tau\right\}$$
$$\le \left|E\left(\rho\left(Y, X^T\beta\right) - \rho\left(Y, X^T\theta_0\right)\right) 1\left\{\tau_0 < Q \le \tau\right\}\right| + \left|E\left(\rho\left(Y, X^T\beta\right) - \rho\left(Y, X^T\beta_0\right)\right) 1\left\{\tau_0 < Q \le \tau\right\}\right|. \tag{A.6}$$

We first look at the second term. For some $\varepsilon > 0$ to be determined, it follows from Assumption 2.2(ii) that there is $C_\varepsilon$,

$$\inf_{E(X^T(\beta-\beta_0))^2 1\{Q \le \tau_0\} > \varepsilon^2} E\left(\rho\left(Y, X^T\beta\right) - \rho\left(Y, X^T\beta_0\right)\right) 1\left\{Q \le \tau_0\right\} > C_\varepsilon.$$

Therefore, if $R(\alpha, \tau) < \min\{C_\varepsilon/2, \varepsilon\}$, we have

$$|E\left(\rho\left(Y, X^T\beta\right) - \rho\left(Y, X^T\beta_0\right)\right) 1\left\{Q \le \tau_0\right\}| \le R(\alpha, \tau) < C_\varepsilon/2$$

because all the three components of $R(\alpha, \tau)$ in (A.5) are non-negative. It implies

$$(E|X^T(\beta - \beta_0)|1\{Q \le \tau_0\})^2 \le E(X^T(\beta - \beta_0)1\{Q \le \tau_0\})^2 < \varepsilon^2$$

In addition, Assumption 2.4 implies that, with $C_1' = C_1^{-1}\left(1 - C_1\right) > 0$ and $C_2' = C_2^{-1}\left(1 - C_2\right) > 0$,

$$C_2' E|X^T\beta| 1\left\{Q > \tau_0\right\} \le E|X^T\beta| 1\left\{Q < \tau_0\right\} \le C_1' E|X^T\beta| 1\left\{Q > \tau_0\right\},$$

for all $\beta \in \mathbb{R}^p$. From the Lipschitz condition,

$$\begin{aligned} & E\left(\rho\left(Y, X^T\beta\right) - \rho\left(Y, X^T\beta_0\right)\right) 1\left\{\tau_0 < Q \le \tau\right\} \le LE\left|X^T\left(\beta - \beta_0\right)\right| 1\left\{\tau_0 < Q \le \tau\right\} \\ & \le LE\left|X^T\left(\beta - \beta_0\right)\right| 1\left\{\tau_0 < Q\right\} \le LC_2'^{-1} E\left|X^T\left(\beta - \beta_0\right)\right| 1\left\{Q \le \tau_0\right\} < LC_2'^{-1}\varepsilon. \end{aligned}$$

On the other hand, the first term on the RHS of (A.6) is the third term in the RHS of (A.5), hence is bounded by $R(\alpha, \tau) < \varepsilon$.

For any $\epsilon' > 0$, it follows from Assumption 2.3 that there is $c > 0$ such that if $\tau > \tau_0 + \epsilon'$. Then

$$\begin{aligned} cP(\tau_0 \ &< \ Q \le \tau_0 + \epsilon') \le cP(\tau_0 < Q \le \tau) \\ &< \ E\left(\rho\left(Y, X^T\beta_0\right) - \rho\left(Y, X^T\theta_0\right)\right) 1\left\{\tau_0 < Q \le \tau\right\} < (LC_2'^{-1} + 1)\varepsilon. \end{aligned}$$

Choosing $\varepsilon = cP(\tau_0 < Q \le \tau_0 + \epsilon')/(2(LC_2'^{-1} + 1))$, we then infer that for this $\varepsilon$, when $R(\alpha, \tau) < \min\{C_\varepsilon/2, \varepsilon\}$, we must have $\tau < \tau_0 + \epsilon'$.

By the same argument, if $\tau < \tau_0$, then we must have $\tau > \tau_0 - \epsilon'$. Hence, $R(\alpha, \tau) < \min\{C_\varepsilon/2, \varepsilon\}$ implies $|\tau - \tau_0| < \epsilon'$.

Now consider the event $\{R(\widehat{\alpha}, \widehat{\tau}) < \min\{C_\varepsilon/2, \varepsilon\}\}$, which occurs with probability approaching one due to Theorem 2.1. On this event, $|\widehat{\tau} - \tau_0| < \epsilon'$. Because $\epsilon'$ is taken arbitrarily, we have proved the consistency of $\widehat{\tau}$.

### A.3 Proof of Theorem 2.3

The proof consists of several steps. First, we prove that $\widehat{\beta}$ is inside the neighborhood of $\beta_0$ so that we can apply the local conditional margin condition of Assumption 2.6. Secondly, we get an intermediate convergence rate for $\widehat{\tau}$ based on the consistency of the risk and $\widehat{\tau}$ assisted by the additional regularity conditions in Section 2.3. Finally, the consistency of $\widehat{\tau}$ enables us to apply assumptions in Section 2.3 to get a tighter bound of the second term in (A.6).

**Step 1** Prove that for any $r > 0$, with probability approaching one (w.p.a.1), if $\widehat{\tau} > \tau_0$, $\widehat{\beta} \in \mathcal{B}(\beta_0, r)$; if $\widehat{\tau} \le \tau_0$, $\widehat{\theta} \in \mathcal{G}(\theta_0, r)$.

For any $r > 0$, by Assumption 2.2, there is $C > 0$, for any $\beta \in \mathbb{R}^p$, if $E(X^T(\beta - \beta_0))^2 1\{Q \le \tau_0\} \ge r^2$, then $E\rho(Y, X^T\beta) - \rho(Y, X^T\beta_0) 1\{Q \le \tau_0\} > C$. Now $R(\widehat{\alpha}, \widehat{\tau}) = o_p(1)$ implies, for the above $C$, the event $R(\widehat{\alpha}, \widehat{\tau}) < C$ holds w.p.a.1. Conditional on this event, by (2.5), when $\beta = \widehat{\beta}$,

$$E(\rho(Y, X^T\beta) - \rho(Y, X^T\beta_0)) 1\{Q \le \tau_0\} \le R(\widehat{\alpha}, \widehat{\tau}) < C.$$

This implies, w.p.a.1, $E(X^T(\beta - \beta_0))^2 1\{Q \le \tau_0\} < r^2$ for $\beta = \widehat{\beta}$ when $\widehat{\tau} > \tau_0$, which is $\widehat{\beta} \in \mathcal{B}(\beta_0, r)$. The case of $\widehat{\tau} \le \tau_0$ can be proved using the same argument, hence is omitted to avoid repetitions.

**Step 2** Show that $|\widehat{\tau} - \tau_0| \le R\left(\widehat{\alpha}, \widehat{\tau}\right)$ w.p.a.1, and get an intermediate convergence rate for $\widehat{\tau}$.

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 $\delta$, apply the Lipschitz continuity, for some $L$ which does not depend on $\beta$ and $\tau$, $\left|\mathrm{E}\left(\rho\left(Y, X^T\beta\right) - \rho\left(Y, X^T\beta_0\right)\right) 1\left\{\tau_0 < Q \le \tau\right\}\right| \le L\mathrm{E}\left|X^T\left(\beta - \beta_0\right)\right| 1\left\{\tau_0 < Q \le \tau\right\}$. Then Assumption 2.5 implies, there is $M > 0$,

$$\begin{aligned}
&\left|\mathrm{E}\left(\rho\left(Y, X^T\beta\right) - \rho\left(Y, X^T\beta_0\right)\right) 1\left\{\tau_0 < Q \le \tau\right\}\right| \\
\text{(Lipschitz)} &\le L\mathrm{E}\left|X^T\left(\beta - \beta_0\right)\right| 1\left\{\tau_0 < Q \le \tau\right\} \\
\text{(Assumption 2.5 (ii))} &\le \left(\tau - \tau_0\right) ML\mathrm{E}\left|X^T\left(\beta - \beta_0\right)\right| 1\left\{\tau_0 < Q\right\} \\
\text{(Assumption 2.4)} &\le \left(\tau - \tau_0\right) MLC_2\mathrm{E}\left|X^T\left(\beta - \beta_0\right)\right| 1\left\{Q \le \tau_0\right\} \\
&\le \left(\tau - \tau_0\right) MLC_2 E\left(X^T\left(\beta - \beta_0\right)\right)^2 1\left\{Q \le \tau_0\right\}. \\
&\le \left(\left(\tau - \tau_0\right) MLC_2\right)^2 / (4c) + cE\left(X^T\left(\beta - \beta_0\right)\right)^2 1\left\{Q \le \tau_0\right\}, \\
\text{(Assumption 2.6)} &\le \left(\left(\tau - \tau_0\right) MLC_2\right)^2 / (4c) + E\left(\rho\left(Y, X^T\beta\right) - \rho\left(Y, X^T\beta_0\right)\right) 1\left\{Q \le \tau_0\right\} \\
\text{(2.5)} &\le \left(\left(\tau - \tau_0\right) MLC_2\right)^2 / (4c) + R(\alpha, \tau),
\end{aligned} \tag{A.7}$$

where the third last inequality follows from $uv \le v^2/(4c) + cu^2$ for any $c > 0$. Note that $M, L, C_2$ are independent of $\tau$. In addition, by (A.6), (2.8) and (2.5),

$$\begin{aligned}
&\left|E\left(\rho\left(Y, X^T\beta\right) - \rho\left(Y, X^T\beta_0\right)\right) 1\left\{\tau_0 < Q \le \tau\right\}\right| \ge E\left(\rho\left(Y, X^T\beta_0\right) - \rho\left(Y, X^T\theta_0\right)\right) 1\left\{\tau_0 < Q \le \tau\right\} \\
&- \left|E\left(\rho\left(Y, X^T\beta\right) - \rho\left(Y, X^T\theta_0\right)\right) 1\left\{\tau_0 < Q \le \tau\right\}\right| \ge c_0(\tau - \tau_0) - R(\alpha, \tau).
\end{aligned}$$

Hence $c_0(\tau - \tau_0) \le \left(\left(\tau - \tau_0\right) MLC_2\right)^2 / (4c) + 2R(\alpha, \tau)$, implying, when $\tau_0 < \tau < \tau_0 + 2cc_0/(MLC_2)^2$, $\tau - \tau_0 \le \frac{4}{c_0} R(\alpha, \tau)$. By the same argument, when $\tau_0 - 2cc_0/(MLC_2)^2 < \tau \le \tau_0$, we have $\tau_0 - \tau \le \frac{4}{c_0} R(\alpha, \tau)$ for $\alpha = (\beta, \delta)$, with any $\theta \in \mathcal{G}(\theta_0, r)$ and arbitrary $\beta$.

Hence when $\widehat{\tau} > \tau_0$, on the event $\widehat{\beta} \in \mathcal{B}(\beta_0, r)$, and $\widehat{\tau} - \tau_0 < 2cc_0/(MLC_2)^2$, we have $\widehat{\tau} - \tau_0 \le \frac{4}{c_0} R(\widehat{\alpha}, \widehat{\tau})$. When $\widehat{\tau} \le \tau_0$, on the event $\widehat{\theta} \in \mathcal{G}(\theta_0, r)$, and $\tau_0 - \widehat{\tau} < 2cc_0/(MLC_2)^2$, we have $\tau_0 - \widehat{\tau} \le \frac{4}{c_0} R(\widehat{\alpha}, \widehat{\tau})$. Hence due to Step 1 and the consistency of $\widehat{\tau}$, we have, w.p.a.1,

$$|\widehat{\tau} - \tau_0| \le R\left(\widehat{\alpha}, \widehat{\tau}\right) = O_p\left(\lambda_n s\right). \tag{A.8}$$

Now applying Lemma A.1 with $a_n$ and $b_n$ replaced by $a_n/2$ and $b_n/2$ and defining

$$\nu_{1n}\left(\tau\right) = \nu_n\left(\alpha_0, \tau\right) - \nu_n\left(\alpha_0, \tau_0\right),$$

we can rewrite the basic inequality in (A.4) by

$$\lambda_n \left|D_0\alpha_0\right|_1 \ge R\left(\widehat{\alpha}, \widehat{\tau}\right) + \lambda_n \left|\widehat{D}\widehat{\alpha}\right|_1 - \frac{1}{2}\lambda_n \left|\widehat{D}\left(\widehat{\alpha} - \alpha_0\right)\right|_1 - \left|\nu_{1n}\left(\widehat{\tau}\right)\right|,$$

which implies that

$$\begin{aligned}&\lambda_n \left( |D_0\alpha_0|_1 - \left|\widehat{D}\alpha_0\right|_1 \right) + |\nu_{1n}(\hat{\tau})| + 2\lambda_n \left|\widehat{D}\left(\widehat{\alpha}-\alpha_0\right)_J\right|_1 \\ &\geq R\left(\widehat{\alpha},\hat{\tau}\right) + \frac{1}{2}\lambda_n\left|\widehat{D}\left(\widehat{\alpha}-\alpha_0\right)\right|_1 .\end{aligned} \tag{A.9}$$

Then, let $c_\alpha = \lambda_n \left( |D_0\alpha_0|_1 - \left|\widehat{D}\alpha_0\right|_1 \right) + |\nu_{1n}(\hat{\tau})|$ and consider two cases: (i) $\lambda_n \left|\widehat{D}\left(\widehat{\alpha}-\alpha_0\right)_J\right| \leq c_\alpha$; (ii) $\lambda_n \left|\widehat{D}\left(\widehat{\alpha}-\alpha_0\right)_J\right| > c_\alpha$.

**Step 3** We begin with Case (ii).

By $\lambda_n \left|\widehat{D}\left(\widehat{\alpha}-\alpha_0\right)_J\right| > c_\alpha$ and the basic inequality (A.9)

$$6\left|\widehat{D}\left(\widehat{\alpha}-\alpha_0\right)_J\right|_1 \geq \left|\widehat{D}\left(\widehat{\alpha}-\alpha_0\right)\right|_1 = \left|\widehat{D}\left(\widehat{\alpha}-\alpha_0\right)_J\right| + \left|\widehat{D}\left(\widehat{\alpha}-\alpha_0\right)_{J^c}\right|, \tag{A.10}$$

which enables us to apply the compatibility condition in Assumption 2.7.

Recall that $\|Z\|_2 = (EZ^2)^{1/2}$ for a random variable $Z$. Using the compatibility coefficient $\phi$ and the inequality $uv \leq v^2/(2c) + cu^2/2$, for $s = |J|$,

$$3\lambda_n \left|\widehat{D}\left(\widehat{\alpha}-\alpha_0\right)_J\right|_1 \leq 3\lambda_n \bar{D} \left\|X(\hat{\tau})^T(\widehat{\alpha}-\alpha_0)\right\|_2 \sqrt{s}/\phi \leq \frac{9\lambda_n^2\bar{D}^2 s}{2c\phi^2} + \frac{c}{2}\left\|X(\hat{\tau})^T(\widehat{\alpha}-\alpha_0)\right\|_2^2, \tag{A.11}$$

**Step 3-I** We now show that for any $r > 0$, w.p.a.1, $\widehat{\beta} \in \mathcal{B}(\beta_0, r)$ and $\widehat{\theta} \in \mathcal{G}(\theta_0, r)$.

Suppose $\widehat{\tau} > \tau_0$, in step 1, we have shown that $\widehat{\beta} \in \mathcal{B}(\beta_0, r)$. We now show that $\widehat{\theta} \in \mathcal{G}(\theta_0, r)$. In fact, by Assumption 2.2, there is $C > 0$, for any $\theta \in \mathbb{R}^p$, if $E(X^T(\theta-\theta_0))^2 1\{Q > \tau_0\} \geq r^2$, then $E\rho(Y, X^T\theta) - \rho(Y, X^T\theta_0)1\{Q > \tau_0\} > C$. Conditional on the event $R(\widehat{\alpha}, \widehat{\tau}) < C/2$, by (2.5), when $\beta = \widehat{\beta}$, and $\tau = \widehat{\tau}$,

$$E(\rho(Y, X^T\theta) - \rho(Y, X^T\theta_0))1\{Q > \tau\} \leq R(\widehat{\alpha}, \widehat{\tau}) < C/2. \tag{A.12}$$

Moreover, by the Lipschitz continuity, and Assumption 2.5,

$$\begin{aligned}&E(\rho(Y, X^T\theta) - \rho(Y, X^T\theta_0))1\{\tau_0 < Q \leq \tau\} \leq LE|X^T(\theta-\theta_0)|1\{\tau_0 < Q \leq \tau\} \\ &\leq L|\theta-\theta_0|_1 E\max_{j\leq p}|X_j|1\{\tau_0 < Q \leq \tau\} \leq L|\theta-\theta_0|_1 E\max_{j\leq p}|X_j|\sup_x P(\tau_0 < Q \leq \tau | X = x) \\ &\leq LM(\tau-\tau_0)|\theta-\theta_0|_1 E\max_{j\leq p}|X_j|.\end{aligned}$$

By the expectation-form of the Bernstein inequality (Lemma 14.12 of Bulmann and van de Geer 2010), $E\max_{j\leq p}|X_j| \leq K_1\log(p+1) + \sqrt{2\log(p+1)}$. By (A.10), $|\widehat{\theta}-\theta_0|_1 = O_p(s)$. Hence by (A.8), $|\widehat{\tau}-\tau_0||\widehat{\theta}-\theta_0|_1 E\max_{j\leq p}|X_j| = O_p(\lambda_n s^2\log p) = o_p(1)$. This implies, w.p.a.1, when $\beta = \widehat{\beta}$, and $\tau = \widehat{\tau}$, $E(\rho(Y, X^T\theta) - \rho(Y, X^T\theta_0))1\{\tau_0 < Q \leq \tau\} < C/2$, and thus with (A.12),

$$E(\rho(Y, X^T\theta) - \rho(Y, X^T\theta_0))1\{\tau_0 < Q\} < C.$$

Hence by Assumption 2.2, $E(X^T(\theta-\theta_0))^2 1\{Q > \tau_0\} < r^2$ for $\theta = \widehat{\theta}$, implying $\widehat{\theta} \in \mathcal{G}(\theta_0, r)$. Therefore, when $\widehat{\tau} > \tau_0$, w.p.a.1, $\widehat{\theta} \in \mathcal{G}(\theta_0, r)$ and $\widehat{\beta} \in \mathcal{B}(\beta_0, r)$. 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 $\widehat{\tau} \leq \tau_0$.

**Step 3-II** Deriving the rate of convergence in case (ii).

Note that, for $\tau > \tau_0$,

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

We bound the three terms on the right hand side when $\alpha = \widehat{\alpha}$ one by one. By Assumption 2.5, there is $C_1 > 0$, when $\beta \in \mathcal{B}(\delta_0, r)$ and $\theta \in \mathcal{G}(\theta_0, r)$, $\delta = \theta - \beta$ satisfies:

$$2\left\|X(\tau)^T\alpha - X(\tau_0)^T\alpha\right\|_2^2 = 2E(X^T\delta)^2 1\{\tau_0 \leq Q < \tau\} \leq C_1(\tau - \tau_0).$$

By (2.3) and Assumption 2.6 (conditional margin condition), with the Lipschitz constant $L > 0$,

$$\begin{aligned} & 4\left\|X(\tau_0)^T\alpha - X(\tau_0)^T\alpha_0\right\|_2^2 = 4E(X^T(\theta-\theta_0))^2 1\{Q \geq \tau_0\} + 4E(X^T(\beta-\beta_0))^2 1\{Q < \tau_0\} \\ \leq\ & C_2 E\left(\rho\left(Y, X^T\theta\right) - \rho\left(Y, X^T\theta_0\right)\right) 1\left\{Q > \tau_0\right\} + C_2 E\left(\rho\left(Y, X^T\beta\right) - \rho\left(Y, X^T\beta_0\right)\right) 1\left\{Q \leq \tau_0\right\} \\ \leq\ & C_2 R(\alpha, \tau) + C_2 E\left(\rho\left(Y, X^T\theta\right) - \rho\left(Y, X^T\theta_0\right)\right) 1\left\{\tau_0 < Q < \tau\right\} \\ \leq\ & C_2 R(\alpha, \tau) + C_2 L E|X^T(\theta-\theta_0)| 1\left\{\tau_0 < Q < \tau\right\} \\ \leq\ & C_2 R(\alpha, \tau) + C_3(\tau - \tau_0). \end{aligned}$$

Moreover, $4\left\|X(\tau_0)^T\alpha_0 - X(\tau)^T\alpha_0\right\|_2^2 = 2E(X^T\delta_0)^2 1\{\tau_0 \leq Q < \tau\} \leq C_4(\tau - \tau_0)$. Therefore, $\left\|X(\tau)^T(\alpha-\alpha_0)\right\|_2^2 \leq C_2 R(\alpha, \tau) + C_5(\tau - \tau_0)$. The case of $\tau \leq \tau_0$ can be proved using the same argument. Hence we have, setting $\tau = \widehat{\tau}$, and $\alpha = \widehat{\alpha}$,

$$\left\|X(\widehat{\tau})^T(\widehat{\alpha}-\alpha_0)\right\|_2^2 \leq C_2 R(\widehat{\alpha}, \widehat{\tau}) + C_5|\widehat{\tau} - \tau_0|.$$

In addition, since $C_5|\widehat{\tau} - \tau_0| \leq C_6 R(\widehat{\alpha}, \widehat{\tau})$, By (A.9), (A.10) and set $c = (C_2 + C_6)^{-1}$ in (A.11),

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

By this and (A.8),

$$|\widehat{\tau} - \tau_0| = O_p(\lambda_n^2 s).$$

**Step 4** We consider Case (i) now.

Due to (A.8) and (A.3) in Lemma A.1, $|\nu_{1n}(\hat{\tau})| = O_p\left(\lambda_n^2 s\right)$ and by the mean value theorem

$$\begin{aligned}
& \lambda_n \left| |D_0\alpha_0|_1 - \left|\widehat{D}\alpha_0\right|_1 \right| \\
\leq\ & \lambda_n \sum_{j=1}^{p} \left( \frac{2}{n} \sum_{i=1}^{n} \left| X_i^{(j)} 1\left\{Q_i > t_0\right\} \right|^2 \right)^{1/2} \left|\delta^{(j)}\right| \frac{1}{n} \sum_{i=1}^{n} \left| X_i^{(j)} \right|^2 \left| 1\left\{Q_i > \widehat{\tau}\right\} - 1\left\{Q_i > \tau_0\right\} \right| \\
=\ & O_p\left(\lambda_n^2 s^2\right). \qquad \text{(A.13)}
\end{aligned}$$

Thus, under Case (i) and by (A.8),

$$\begin{aligned}
c_0 \left|\widehat{\tau} - \tau_0\right| &\leq \frac{\lambda_n}{2} \left|\widehat{D}\left(\widehat{\alpha} - \alpha_0\right)\right|_1 + R\left(\widehat{\alpha}, \widehat{\tau}\right) \\
&\leq 3\lambda_n \left( |D_0\alpha_0|_1 - \left|\widehat{D}\alpha_0\right|_1 \right) + 3\left|\nu_{1n}\left(\widehat{\tau}\right)\right| \\
&\leq O_p\left(\lambda_n^2 s^2\right) + O_p\left(\lambda_n^2 s\right). \qquad \text{(A.14)}
\end{aligned}$$

Note that we can plug in (A.14) into (A.13) to update the bound as $O_p\left(\lambda_n^3 s^3\right)$, which then updates the bound in (A.14) as $O_p\left(\lambda_n^3 s^3\right) + O_p\left(\lambda_n^2 s\right)$. Repeat this $k$ times so that $\lambda_n^{k+1} s^{k+1} \leq \lambda_n^2 s$, which exists since $\lambda_n s^2 \to 0$, and that

$$\left|\widehat{\tau} - \tau_0\right| \leq O_p\left(\lambda_n^{k+1} s^{k+1}\right) + O_p\left(\lambda_n^2 s\right) = O_p\left(\lambda_n^2 s\right),$$

as well as $\left|\widehat{D}\left(\widehat{\alpha} - \alpha_0\right)_J\right| = O_p\left(\lambda_n s\right)$. Combining cases 1 and 2, because $D(\widehat{\tau}) \geq \underline{D}$ we have

$$|\widehat{\tau} - \tau_0| = O_p(\lambda_n^2 s), \quad |\widehat{\alpha} - \alpha_0|_1 = O_p(\lambda_n s).$$

This completes the proof.

# B Proofs for Section 3

## B.1 Proof of Theorem 3.1

Define the oracle space $S = \{\alpha \in \mathbb{R}^{2p} : \alpha_j = 0, j \in J\}$ and use the convention that $\widetilde{Q}_n(\alpha_J) = \widetilde{Q}_n(\alpha_J, 0)$ so that

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

Throughout this section, our argument is conditional on

$$\widehat{\tau} \in \mathcal{T}_n = \left\{ |\tau - \tau_0| \leq \omega_n^2 |J| \cdot \log n \right\}, \qquad \text{(B.1)}$$

whose probability goes to 1 due to Theorem 2.3.

The proof consists of the following two lemmas.

**Lemma B.1.** *Suppose $s^4 \log s = o(n)$. Let $\widetilde{\alpha}_J = \operatorname{argmin}_{\alpha_J} \widetilde{Q}_n(\alpha_J)$, then*

$$|\widetilde{\alpha}_J - \alpha_{0J}|_2 = O_p(\sqrt{\frac{s \log s}{n}}).$$

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

$$\inf_{|\alpha_J - \alpha_{0J}|_2 = C_\delta k_n} \widetilde{Q}_n(\alpha_J) > \widetilde{Q}_n(\alpha_{0J}) \tag{B.2}$$

Once this is proved, then by the convexity of $\widetilde{Q}_n$, there is a local minimizer of $\widetilde{Q}_n(\alpha_J)$ inside $B(\alpha_{0J}, C_\delta k_n) = \{\alpha_J \in \mathbb{R}^s : |\alpha_{0J} - \alpha_J|_2 \leq C_\delta k_n\}$. We now prove (B.2), conditioning on the event $\widehat{\tau} \in \mathcal{T}_n$ in (B.1), whose probability goes to one.

Write

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

Then

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

To analyze (1), note that $|\alpha_J - \alpha_{0J}|_2 = C_\delta k_n$ and $m_J(\tau_0, \alpha_0) = 0$. On the event $\widehat{\tau} \in \mathcal{T}_n$, there is $c_3 > 0$,

$$\begin{aligned}
& L_J(\widehat{\tau}, \alpha_J) - L_J(\widehat{\tau}, \alpha_{0J}) \\
\geq \quad & m_J(\tau_0, \alpha_{0J})^T (\alpha_J - \alpha_{0J}) + (\alpha_J - \alpha_{0J})^T \frac{\partial^2 E\rho(Y, X_J(\widehat{\tau})^T \alpha_{0J})}{\partial \alpha_J \partial \alpha_J^T} (\alpha_J - \alpha_{0J}) \\
& -|m_J(\tau_0, \alpha_{0J}) - m_J(\widehat{\tau}, \alpha_{0J})|_2 |\alpha_J - \alpha_{0J}|_2 - c_3 |\alpha_{0J} - \alpha_J|_1^3 \\
\geq \quad & \lambda_{\min}(\frac{\partial^2 E\rho(Y, X_J(\widehat{\tau})^T \alpha_{0J})}{\partial \alpha_J \partial \alpha_J^T}) |\alpha_J - \alpha_{0J}|_2^2 \\
& -(|m_J(\tau_0, \alpha_{0J}) - m_J(\widehat{\tau}, \alpha_{0J})|_2) |\alpha_J - \alpha_{0J}|_2 - c_3 s^{3/2} |\alpha_{0J} - \alpha_J|_2^3 \\
\geq \quad & c_2 C_\delta^2 k_n^2 - (|m_J(\tau_0, \alpha_{0J}) - m_J(\widehat{\tau}, \alpha_{0J})|_2) C_\delta k_n - c_3 s^{3/2} C_\delta^3 k_n^3 \\
\\
\geq \quad & C_\delta k_n (c_2 C_\delta k_n - M(s) \lambda_n^2 |J| \cdot \log n - c_s s^{3/2} C_\delta^2 k_n^2) \geq c_2 C_\delta^2 k_n^2 / 3,
\end{aligned}$$

where the last inequality holds for all large $n$ since $s^{3/2} k_n = o(1)$ and $c_2 C_\delta k_n > 3M(s) \lambda_n^2 |J| \cdot \log n$.

To analyze (2), by the Symmetrization theorem (van der Vaart and Wellner 1996) and contraction theorem (Theorem 14.4 of Bulhmann and van der Geer 2009), there is a Rademacher sequence

$\epsilon_1, ..., \epsilon_n$ independent of $\{Y_i, X_i, Q_i\}_{i\leq n}$ so that

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

which is bounded by the sum of the following two terms due to the triangle inequality and the fact that $|\alpha_J - \alpha_{0J}|_1 \leq |\alpha_J - \alpha_{0J}|_2\sqrt{s}$,

$$\begin{aligned}
V_{1n} &= 2LE \sup_{\tau\in\mathcal{T}_n} \sup_{|\alpha_J-\alpha_{0J}|_1\leq C_\delta k_n\sqrt{s}} |\frac{1}{n}\sum_{i=1}^n \epsilon_i(X_{iJ}(\tau) - X_{iJ}(\tau_0))^T(\alpha_J - \alpha_{0J})| \\
&\leq 2LE \sup_{\tau\in\mathcal{T}_n} \sup_{|\delta_{J_2}-\delta_{0J_2}|_1\leq C_\delta k_n\sqrt{s}} |\frac{1}{n}\sum_{i=1}^n \epsilon_i X_{iJ_2}^T(1\{Q_i > \tau\} - 1\{Q_i > \tau_0\})(\delta_{J_2} - \delta_{0J_2})| \\
&\leq 2LC_\delta k_n\sqrt{s}E \sup_{\tau\in\mathcal{T}_n} \max_{j\in J_2} \left|\frac{1}{n}\sum_{i=1}^n \epsilon_i X_{ij}(1\{Q_i > \tau\} - 1\{Q_i > \tau_0\})\right| \\
&\leq 2LC_\delta k_n\sqrt{s}C_1\,|J_2|\sqrt{\frac{\omega_n^2|J|\cdot\log n}{n}},
\end{aligned}$$

due to the maximal inequality (for VC class), and

$$\begin{aligned}
V_{2n} &= 2LE \sup_{|\alpha_J-\alpha_{0J}|_1\leq C_\delta k_n\sqrt{s}} |\frac{1}{n}\sum_{i=1}^n \epsilon_i X_{iJ}(\tau_0)^T(\alpha_J - \alpha_{0J})| \\
&\leq E \max_{j\in J} |\frac{1}{n}\sum_{i=1}^n \epsilon_i X_{ij}(\tau_0)| \leq 2LC_\delta k_n\sqrt{s}C_2\sqrt{\frac{\log s}{n}},
\end{aligned}$$

due to the Bernstein's moment inequality (Lemma 14.12 of Bulhmann and van der Geer 2010) for some $C_2 > 0$.

Recall the definition of $\omega_n$ in (2.10) and note that $C_1\,|J_2|\,\omega_n\sqrt{\frac{|J|\cdot\log n}{n}} < C_2\sqrt{\frac{\log s}{n}}$ for all sufficiently large $n$. Hence

$$V_n \leq 4LC_\delta k_n\sqrt{s}C_2\sqrt{\frac{\log s}{n}}$$

Therefore, conditioning on the event $\widehat{\tau}\in\mathcal{T}_n$, with probability at least $1-\delta$, (2) $\leq \frac{1}{\delta}4LC_2C_\delta k_n^2$.

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

$$\inf_{|\alpha_J-\alpha_{0J}|_2=C_\delta k_n} \widetilde{Q}_n(\alpha_J) - \widetilde{Q}_n(\alpha_{0J}) \geq \frac{c_2C_\delta^2k_n^2}{3} - \frac{1}{\delta}4LC_2C_\delta k_n^2 > 0.$$

The last inequality holds for $C_\delta > \frac{12LC_2}{c_2\delta}$. By the convexity of $\widetilde{Q}_n$, there is a local minimizer of $\widetilde{Q}_n(\alpha_J)$ inside $\{\alpha_J \in \mathbb{R}^s : |\alpha_{0J} - \alpha_J|_2 \leq C_\delta k_n\}$. Q.E.D.

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 $\widetilde{\alpha}_J=(\widetilde{\beta}_{J_1},\widetilde{\delta}_{J_2})$ in the previous lemma, define

$$\widetilde{\alpha}=(\widetilde{\beta}_{J_1},0,\widetilde{\delta}_{J_2},0)$$

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

**Lemma B.2.** *With probability approaching one, there is a random neighborhood of $\widetilde{\alpha}$ 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 $\widetilde{Q}_n(\alpha_J,0)<\widetilde{Q}_n(\alpha)$.*

Define an $L_2$- ball, for $r_n=\mu_n/\log n$,

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

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

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

Note that $|(\alpha_J,0)-\widetilde{\alpha}|_2^2=|\alpha_J-\widetilde{\alpha}_J|_2^2\leq|\alpha_J-\widetilde{\alpha}_J|_2^2+|\alpha_{J^c}-0|_2^2=|\alpha-\widetilde{\alpha}|_2^2$. Hence $\alpha\in\mathcal{H}$ implies $(\alpha_J,0)\in\mathcal{H}$. In addition, by definition of $\widetilde{\alpha}=(\widetilde{\alpha}_J,0)$ and $|\widetilde{\alpha}_J-\alpha_{0J}|_2=O_p(\sqrt{\frac{s\log s}{n}})$, we have $|\widetilde{\alpha}-\alpha_0|_1=O_p(s\sqrt{\frac{\log s}{n}})$, 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 $\alpha$ and $(\alpha_J,0)$,

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

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

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

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

By Assumption 3.1, $\sup_{\alpha \in \mathcal{H}} |\alpha - \alpha_0|_1 = O_p(s\sqrt{\frac{\log s}{n}}) + r_n$ implies, there is $L > 0$, for any $\delta > 0$, there is $C_\delta > 0$, with probability at last $1 - \delta$,

$$\max_{j \notin J} \sup_{\tau \in \mathcal{T}_n} \sup_{\alpha \in \mathcal{H}} |m_j(\tau, \alpha) - m_j(\tau, \alpha_0)| \leq L \sup_{\alpha \in \mathcal{H}} |\alpha - \alpha_0|_1 \leq L(C_\delta s\sqrt{\frac{\log s}{n}} + r_n),$$
$$\max_{j \leq 2p} \sup_{\tau \in \mathcal{T}_n} |m_j(\tau, \alpha_0) - m_j(\tau_0, \alpha_0)| \leq M_n|\tau - \tau_0| \leq M_n\omega_n^2|J| \cdot \log n.$$

Therefore, $\sup_{j \notin J} \sup_{\tau \in \mathcal{T}_n} |m_j(\tau, h)| = O_p(s\sqrt{\frac{\log s}{n}} + r_n + M_n\omega_n^2|J| \cdot \log n) = o_p(\mu_n)$.

On the other hand, from the trianglue inequality, Lemma A.1 $(A.1)$ and the fact that $a_n < \lambda_n$, it follows that

$$\begin{aligned} E \sup_{|\alpha - \alpha_0| \leq c_n} \sup_{\tau \in \mathcal{T}_n} |\nu_n(\tau, \alpha_J, 0) - \nu_n(\tau, \alpha)| &\leq 2E \sup_{|\alpha - \alpha_0| \leq c_n} \sup_{\tau \in \mathcal{T}_n} |\nu_n(\tau, \alpha) - \nu_n(\tau, \alpha_0)| \\ &\leq 2\lambda_n c_n = o\left(\mu_n c_n\right), \end{aligned}$$

where $c_n = Cs\sqrt{(\log s)/n} + r_n$. Combining these, $L_n(\tau, \alpha_J, 0) - L_n(\tau, \alpha) = o_p(\mu_n) \sum_{j \notin J} |\alpha_j|$.

On the other hand, $\sum_{j \in J} w_j \mu_n \widehat{D}_j |\alpha_j| - \sum_j w_j \mu_n \widehat{D}_j |\alpha_j| = \sum_{j \notin J} \mu_n w_j \widehat{D}_j |\alpha_j|$. Also, with probability approaching one, $w_j = 1$and $\widehat{D}_j \geq \bar{D}$ for all $j \notin J$. Hence with probability approaching one, $\widetilde{Q}_n(\alpha_J, 0) - \widetilde{Q}_n(\alpha)$ equals

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

Q.E.D.

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

$$\widetilde{Q}_n(\widetilde{\alpha}) \leq \widetilde{Q}_n(\alpha_J, 0) \leq \widetilde{Q}_n(\alpha).$$

This implies that $\widetilde{\alpha}$ is a local minimizer of $\widetilde{Q}_n$ in $\mathcal{H}$. Because $\widetilde{Q}_n$ is convex in $\alpha$, $\widetilde{\alpha}$ is also a global minimizer in $\mathbb{R}^{2p}$. This finishes the proof.

## B.2 Proof of Theorem 3.2

**We assume $\rho(y, t)$ to be twice differentiable in $t$**

**Assumptions**

$$\max_{j,l\in J}\mathrm{var}(\frac{\partial^2\rho(Y,X_J(\tau_0)^T\alpha_{0J})}{\partial\alpha_j\partial\alpha_l}<\infty$$

$$\lambda_{\min}(\mathrm{var}(\frac{\partial}{\partial\alpha_J}\rho(Y_i,X_{iJ}^T(\tau_0)\alpha_{0J}))) > c_1$$

Also, the second derivative is Lipchitz continuous wih respect to $X^T\alpha_J$.

We aim to prove: for any $|v|_2=1$, $v\in\mathbb{R}^J$,

$$\sqrt{n}v^T\Sigma_1^{-1/2}\Sigma_2(\widetilde{\alpha}_J-\alpha_{0J})\to^d N(0,1)$$

where $\Sigma_1=\mathrm{var}(\frac{\partial}{\partial\alpha_J}\rho(Y_i,X_{iJ}^T(\tau_0)\alpha_{0J}))$ and $\Sigma_2=E\frac{\partial^2}{\partial\alpha_J\partial\alpha_J^T}\rho(Y_i,X_{iJ}^T(\tau_0)\alpha_{0J})$.

*Proof.* The first order condition implies, because $P(w_j=0\forall j\in J)\to 1$, with probability approaching one, for some $\alpha_J^*$ on the segment joining $\alpha_{0J}$ and $\widehat{\alpha}_J$,

$$0=\frac{1}{n}\sum_{i=1}^n\frac{\partial}{\partial\alpha_J}\rho(Y_i,X_{iJ}^T(\widehat{\tau})\widetilde{\alpha}_J)=\frac{1}{n}\sum_{i=1}^n\frac{\partial}{\partial\alpha_J}\rho(Y_i,X_{iJ}^T(\widehat{\tau})\alpha_{0J})+\frac{1}{n}\sum_{i=1}^n\frac{\partial^2}{\partial\alpha_J\partial\alpha_J^T}\rho(Y_i,X_{iJ}^T(\widehat{\tau})\alpha_J^*)(\widetilde{\alpha}_J-\alpha_{0J})$$

Define

$$H_n(\tau,\alpha_J)=\frac{1}{n}\sum_{i=1}^n\frac{\partial^2}{\partial\alpha_J\partial\alpha_J^T}\rho(Y_i,X_{iJ}^T(\tau)\alpha_J),\quad H(\tau,\alpha_J)=EH_n(\tau,\alpha_J)$$

then $-\Sigma_1^{-1/2}\frac{1}{\sqrt{n}}\sum_{i=1}^n\frac{\partial}{\partial\alpha_J}\rho(Y_i,X_{iJ}^T(\tau_0)\alpha_{0J})=\sqrt{n}\Sigma_1^{-1/2}H(\tau_0,\alpha_{0J})(\widetilde{\alpha}_J-\alpha_{0J})+\sum_{i=1}^3R_i$, where

$$\begin{aligned}R_1&=\sqrt{n}\Sigma_1^{-1/2}(H_n(\tau_0,\alpha_{0J})-H(\tau_0,\alpha_{0J}))(\widetilde{\alpha}_J-\alpha_{0J})\\R_2&=\sqrt{n}\Sigma_1^{-1/2}(H_n(\widehat{\tau},\alpha_J^*)-H_n(\tau_0,\alpha_{0J}))(\widetilde{\alpha}_J-\alpha_{0J})\\R_3&=\sqrt{n}\Sigma_1^{-1/2}\left(\frac{1}{n}\sum_{i=1}^n\frac{\partial}{\partial\alpha_J}\rho(Y_i,X_{iJ}^T(\widehat{\tau})\alpha_{0J})-\frac{1}{n}\sum_{i=1}^n\frac{\partial}{\partial\alpha_J}\rho(Y_i,X_{iJ}^T(\tau_0)\alpha_{0J})\right).\end{aligned}$$

We now examine $R_i, i=1,2,3$. Note that

$$\begin{aligned}E\|H_n(\tau_0,\alpha_{0J})-H(\tau_0,\alpha_{0J})\|_F^2&=\sum_{j\in J}\sum_{l\in J}E[\frac{1}{n}\sum_{i=1}^n\frac{\partial^2\rho(Y_i,X_{iJ}(\tau_0)^T\alpha_{0J})}{\partial\alpha_j\partial\alpha_l}-E\frac{\partial^2\rho(Y,X_J(\tau_0)^T\alpha_{0J})}{\partial\alpha_j\partial\alpha_l}]^2\\&=\sum_{j\in J}\sum_{l\in J}\frac{1}{n}\mathrm{var}(\frac{\partial^2\rho(Y,X_J(\tau_0)^T\alpha_{0J})}{\partial\alpha_j\partial\alpha_l})\leq C\frac{s^2}{n}\end{aligned}$$

Hence $R_1=O_p(\sqrt{\frac{s\log s}{n}})$. Also,

$$\|H_n(\tau_0,\alpha_{0J})-H_n(\widehat{\tau},\alpha_J^*)\|_F^2=\sum_{j\in J}\sum_{l\in J}[\frac{1}{n}\sum_{i=1}^n\frac{\partial^2\rho(Y_i,X_{iJ}(\tau_0)^T\alpha_{0J})}{\partial\alpha_j\partial\alpha_l}-\frac{\partial^2\rho(Y_{,i}\,X_{iJ}(\widehat{\tau})^T\alpha_J^*)}{\partial\alpha_j\partial\alpha_l}]^2$$

$$\leq\sum_{j\in J}\sum_{l\in J}2[\frac{1}{n}\sum_{i=1}^n\frac{\partial^2\rho(Y_i,X_{iJ}(\tau_0)^T\alpha_{0J})}{\partial\alpha_j\partial\alpha_l}-\frac{\partial^2\rho(Y_{,i}\,X_{iJ}(\tau_0)^T\alpha_J^*)}{\partial\alpha_j\partial\alpha_l}]^2$$

$$+2[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2\rho(Y_i, X_{iJ}(\tau_0)^T\alpha_J^*)}{\partial\alpha_j\partial\alpha_l}-\frac{\partial^2\rho(Y_{,i}\, X_{iJ}(\widehat{\tau})^T\alpha_J^*)}{\partial\alpha_j\partial\alpha_l}]^2=...$$

**This proof is very standard, and is not completely finished. I will finish it later, depending on how do we plan about the non-smooth case**

Hence $R_2 = o_p(1)$. Finally, $R_3 = o_p(1)$. The result then follows from

$$\sqrt{n}v^T\Sigma_1^{-1/2}\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\alpha_J}\rho(Y_i, X_{iJ}^T(\tau_0)\alpha_{0J}) \to^d N(0,1).$$

# C Proofs of Section 4

## C.1 Proof of Lemma 4.1

*Proof.* To verify Assumption 2.6, note that $\rho(Y,t) = h_\gamma(Y-t)$, where $h_\gamma(t) = t(\gamma - 1\{t \le 0\})$. By (B.3) of Belloni and Chernozhukov (2011),

$$h_\gamma(w-v) - h_\gamma(w) = -v(\gamma - 1\{w \le 0\}) + \int_0^v (1\{w \le z\} - 1\{w \le 0\})dz$$

where $w = Y - X^T\beta_0$ and $v = X^T\delta_0$. Note that

$$Ev(\gamma - 1\{w \le 0\})1\{\tau < Q \le \tau_0\} = EX^T\delta_0(\gamma - 1\{U \le 0\})1\{\tau < Q \le \tau_0\} = 0$$

given $P(U \le 0|X,Q) = \gamma$. Hence

$$\begin{aligned}
&\mathrm{E}\left(\rho\left(Y, X^T\theta_0\right) - \rho\left(Y, X^T\beta_0\right)\right)1\left\{\tau < Q_i \le \tau_0\right\}\\
&= E\int_0^{X^T\delta_0}(1\{Y \le X^T\beta_0 + z) - 1\{Y \le X^T\beta_0\})dz1\{\tau < Q < \tau_0\}\\
&= \int_\tau^{\tau_0} E\left\{\int_0^{X^T\delta_0}[F_{Y|X,Q}(X^T\beta_0 + z) - F_{Y|X,Q}(X^T\beta_0)]dz\middle|Q = q\right\}dF_Q(q) \equiv \int_\tau^{\tau_0} M(q)dF_Q(q),
\end{aligned}$$

where $F_{Y|X,Q}$ denotes the CDF of the conditional distribution $Y|X,Q$ and

$$M(q) = E\left\{\int_0^{X^T\delta_0}[F_{Y|X,Q}(X^T\beta_0 + z) - F_{Y|X,Q}(X^T\beta_0)]dz\middle|Q = q\right\}.$$

Then using the mean value expansion, the same argument as (B.4) of Belloni and Chernozhukov (2011) implies: for some $t \in [0,z]$,

$$\begin{aligned}
M(q) &= E\left\{\int_0^{X^T\delta_0}[zf_{Y|X,Q}(X^T\beta_0|X,Q) + \frac{z^2}{2}f'_{Y|X,Q}(X^T\beta_0 + t)]dz\middle|Q = q\right\}\\
&= \frac{1}{2}\delta_0^T E[XX^T f_{Y|X,Q}(X^T\beta_0|X,Q)|Q = q]\delta_0 + E\left\{\int_0^{X^T\delta_0}\frac{z^2}{2}f'_{Y|X,Q}(X^T\beta_0 + t)dz\middle|Q = q\right\}
\end{aligned}$$

Note that $|f'_{Y|X,Q}(X^T\beta_0+t)|<C_1$, hence

$$|E\left\{\int_0^{X^T\delta_0}\frac{z^2}{2}f'_{Y|X,Q}(X^T\beta_0+t)dz\middle|Q\right\}|\leq\frac{C_1}{6}E((X^T\delta_0)^3|Q).$$

Define $H(q)=E(f_{Y|X,Q}(X^T\beta_0|X,Q)XX^T|Q=q)$. Therefore,

$$\begin{aligned}M(q) &\geq \frac{1}{2}\delta_0^TH(q)\delta_0-\frac{C_1}{6}E((X^T\delta_0)^3|Q=q)\\ &\geq \frac{1}{2}C_2\delta_0^TE(XX^T|Q=q)\delta_0-\frac{C_1}{6}E((X^T\delta_0)^3|Q=q)\\ &\geq \frac{1}{4}C_2\delta_0^TE(XX^T|Q=q)\delta_0=\frac{C_2}{4}E((X^T\delta_0)^2|Q=q).\end{aligned}$$

The last inequality follows from condition (iii) of Assumption 4.1. This implies

$$\mathrm{E}\left(\rho\left(Y,X^T\theta_0\right)-\rho\left(Y,X^T\beta_0\right)\right)1\left\{\tau<Q_i\leq\tau_0\right\}\geq\frac{C_2}{4}E((X^T\delta_0)^21\{\tau<Q\leq\tau_0\}$$

Therefore the first inequality of Assumption 2.6 is verified with $G(u)=\frac{C_2}{4}u^2$. All the other three inequalities can be verified using the same argument. Note that the constant $C_2$ in $G(u)$ can be the same for all the four inequalities in this assumption, which is the lower bound of $f_{Y|X,Q}(y|x,q)$.

To verify Assumption 3.1, recall that $m_j(\tau,\alpha)=E[X_j(\tau)(1\{Y-X(\tau)^T\alpha\leq 0\}-\gamma)]$. For condition (i) of Assumption 3.1, for all $j\leq 2p$, note that $m_j(\tau_0,\alpha_0)=0$, for all $j\leq 2p$,

$$\begin{aligned}&|m_j(\tau,\alpha_0)-m_j(\tau_0,\alpha_0)|=|EX_j(\tau)[1\{Y\leq X(\tau)^T\alpha_0\}-1\{Y\leq X(\tau_0)^T\alpha_0\}]|\\ &=|EX_j(\tau)[P(Y\leq X(\tau)^T\alpha_0|X,Q)-P(Y\leq X(\tau_0)^T\alpha_0|X,Q)]|\\ &\leq C_3E|X_j(\tau)||(X(\tau)-X(\tau_0))^T\alpha_0|=C_3E|X_j(\tau)||X^T\delta_0(1\{Q>\tau\}-1\{Q>\tau_0\})|\\ &\leq C_3E|X_j(\tau)X^T\delta_0|(1\{\tau<Q<\tau_0\}+1\{\tau_0<Q<\tau\})\\ &\leq C_3(P(\tau_0<Q<\tau)+P(\tau<Q<\tau_0))\sup_q E(|X_j(\tau)X^T\delta_0||Q=q)=O(s)|\tau_0-\tau|\end{aligned}$$

where the first inequality is due to Assumption 4.1, and the last $O(s)$ is uniformly in $j\leq 2p$.

We now verify Condition (ii) of Assumption 3.1. For all $j$ and $\tau$ in a neighborhood of $\tau_0$,

$$\begin{aligned}&|m_j(\tau,\alpha)-m_j(\tau,\alpha_0)|=|EX_j(\tau)(1\{Y\leq X(\tau)^T\alpha\}-1\{Y\leq X(\tau)^T\alpha_0\})|\\ &=|EX_j(\tau)(P(Y\leq X(\tau)^T\alpha|X,Q)-P(Y\leq X(\tau)^T\alpha_0|X,Q))|\\ &\leq C_3E|X_j(\tau)||X(\tau)^T(\alpha-\alpha_0)|\leq C_3|\alpha-\alpha_0|_1\max_{j\leq 2p,i\leq 2p}E|X_j(\tau)X_i(\tau)|,\end{aligned}$$

which implies the result. Finally, the score condition $m(\tau_0,\alpha_0)=0$ in Condition (iii) holds because $P(U\leq 0|X,Q)=\gamma$, and the rest of condition (iii) are also straightforward to verify.

## C.2 Proof of Lemma 4.2

*Proof.* We first prove that Assumption 2.6 holds for the binary choice model. For the first inequality of Assumption 2.6, define

$$f(z) = -g(X^T\beta_0)\log\frac{z}{1-z} - \log(1-z), \quad z \in (0,1).$$

Then straightforward calculation proves that, for any $\beta \in \mathbb{R}^p$,

$$E\rho(Y, X^T\beta)1\{\tau < Q < \tau_0\} = Ef(g(X^T\beta))1\{\tau < Q < \tau_0\}.$$

Thus $E[\rho(Y,X^T\theta_0) - \rho(Y,X^T\beta_0)]1\{\tau < Q < \tau_0\} = E[f(g(X^T\theta_0)) - f(g(X^T\beta_0))]1\{\tau < Q < \tau_0\}$. Let $z_1 = g(X^T\theta_0)$, $z_2 = g(X^T\beta_0)$. Then $f'(z_2) = 0$ almost surely. By Taylor's expansion, there are $\lambda \in [0,1]$ and $\tilde{z}$ such that $f(z_1) - f(z_2) = \frac{f''(\tilde{z})}{2}(z_1 - z_2)^2$, which implies, for $\beta = \lambda\theta_0 + (1-\lambda)\beta_0$,

$$E[\rho(Y,X^T\theta_0) - \rho(Y,X^T\beta_0)]1\{\tau < Q < \tau_0\} = E[\frac{f''(\tilde{z})}{2}g'(X^T\beta)^2(X^T\theta_0 - X^T\beta_0)^2]1\{\tau < Q < \tau_0\}.$$

In addition, by Assumption 4.2, $f''(z) = \frac{g(X^T\beta_0)}{z^2} + \frac{1-g(X^T\beta_0)}{(1-z)^2} > C_2$ and $g'(X^T\beta)^2 > C_2$. Hence

$$E[\rho(Y,X^T\theta_0) - \rho(Y,X^T\beta_0)]1\{\tau < Q < \tau_0\} \geq \frac{C_2^2}{2}E(X^T\theta_0 - X^T\beta_0)^2 1\{\tau < Q < \tau_0\}.$$

All the other three inequalities can be proved following the same lines of proofs, hence are omitted to avoid repetitions.

For part (i) of Assumption 3.1, by the mean value theorem, for all $j \leq 2p$,

$$\begin{aligned}
&|m_j(\tau,\alpha_0) - m_j(\tau_0,\alpha_0)| = \left|E\left\{\frac{g(X(\tau_0)^T\alpha_0) - g(X(\tau)^T\alpha_0)}{g(X(\tau)^T\alpha_0)(1-g(X(\tau)^T\alpha_0))}g'(X(\tau)^T\alpha)X_j(\tau)\right\}\right| \\
&\leq C\sup_t |g'^2 E|X^T\delta_0(1\{Q > \tau_0\} - 1\{Q > \tau\})X_j(\tau)| \\
&\leq C'(P(\tau_0 < Q < \tau) + P(\tau < Q < \tau_0))\sup_q E(|X_j(\tau)X^T\delta_0||Q = q) = O(s)|\tau_0 - \tau|
\end{aligned}$$

where $O(s)$ is uniform in $j$. For part (ii), by the mean value theorem, there is $Z$,

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

Condition (iii) can be verified by straightforward calcualations.