# Weighted Lasso Bootstrap

Tun Lee Ng Michael A. Newton

August 29, 2018

## 1 Introduction

Consider the following linear regression model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \epsilon_i, \tag{1}$$

for $i = 1, \ldots, n$, and $\{\epsilon_i\}$ are independent and identically distributed (iid) random variables with mean 0 and finite variance $\sigma^2$. We assume that $p$ is fixed. Without loss of generality, the covariates are centered to have mean 0, so that $\hat{\beta}_0 = \bar{Y}$, and $Y_i$ in (1) can be replaced by $Y_i - \bar{Y}$. Again, without loss of generality, we assume that $\bar{Y} = 0$. Then, (1) can be expressed as the following

$$Y_i = \boldsymbol{x}_i' \boldsymbol{\beta} + \epsilon_i, \tag{2}$$

where $Y_i$ is the centered response, $\boldsymbol{x}_i' = (x_{i1}, \ldots, x_{ip})$ is the $p \times 1$ centered covariate vector and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ are the regression parameters.

Throughout this paper, we assume that our predictor matrix $X$ satisfies the following conditions:

$$\max_{1 \leq i \leq n} \|\boldsymbol{x}_i\|_2^2 = \mathcal{O}(1) \quad \text{as } n \to \infty, \tag{3}$$

and there exists a non-singular matrix $C$ such that

$$\frac{1}{n} X'X = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' \to C \quad \text{as } n \to \infty. \tag{4}$$

Let $\boldsymbol{\beta}_0$ be the true values of the regression parameters $\boldsymbol{\beta}$. The model is assumed to be sparse, ie. some of the elements of $\boldsymbol{\beta}_0$ are exactly zero corresponding to predictors that are irrelevant to the response.

The Lasso estimator is defined to be the minimizer of the $l_1$-penalized least square objective function,

$$\widehat{\boldsymbol{\beta}}_n := \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2 + \lambda_n \sum_{j=1}^{p} |\beta_j| \tag{5}$$

for a given penalty or regularization parameter $\lambda_n$. The Lasso estimator was first introduced by Tibshirani (1996). Knight and Fu (2000) obtained the asymptotic distribution of the Lasso estimator and showed that the Lasso is weakly consistent under some mild regularity condition. Chatterjee and Lahiri (2011) studied strong consistency of the Lasso estimator under a slightly more stringent regularity condition.

Following the idea by Newton and Raftery (1994), for a given set of responses $\boldsymbol{y} = (y_1, \ldots, y_n)'$, we define the weighted Lasso estimator as follows:

$$\widehat{\boldsymbol{\beta}}_n^w := \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \widetilde{w}_i (y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2 + \lambda_n \widetilde{w}_{n+1} \sum_{j=1}^{p} |\beta_j|. \tag{6}$$

Here, $\widetilde{\boldsymbol{w}} = (\widetilde{w}_1, \ldots, \widetilde{w}_{n+1})$ are random weights drawn from

$$\left( \frac{W_1}{\sum_{i=1}^{n+1} W_i}, \ldots, \frac{W_{n+1}}{\sum_{i=1}^{n+1} W_i} \right) = \left( \frac{W_1}{(n+1)\overline{W}}, \ldots, \frac{W_{n+1}}{(n+1)\overline{W}} \right),$$

where $W_1, \ldots, W_n \overset{iid}{\sim} \exp(1)$ and $W_{n+1} = 1$ a.s. Note that the random weights $\widetilde{\boldsymbol{w}}$ are generated independently of the data $\boldsymbol{y}$, and are similar in structure to a Dirichlet weight vector as expounded by Newton and Raftery (1994).

Hence, for a given set of data $\boldsymbol{y}$, (6) can be expressed as

$$\widehat{\boldsymbol{\beta}}_n^w = \arg\min_{\boldsymbol{\beta}} \frac{1}{\overline{W}} \left\{ \frac{1}{n} \sum_{i=1}^{n} W_i (y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2 + \frac{\lambda_n}{n} \sum_{j=1}^{p} |\beta_j| \right\}. \tag{7}$$

For any given set of data, the sampling distribution of $\left\{ \widehat{\boldsymbol{\beta}}_{n,k}^w \right\}_{k=1}^{K}$ is induced by the randomly drawn weights $\{\widetilde{\boldsymbol{w}}_k\}_{k=1}^{K}$.

# 2 Asymptotics for WLB

Need to define the proper probability space (Newton, 1991)...
Need to define "*convergence in conditional probability*" (denoted with $\xrightarrow{\text{c.p.}}$)...
Need to define "*convergence in conditional distribution*" (denoted with $\xrightarrow{\text{c.d.}}$)...
Need restrictions on the topology of parameter space ($\boldsymbol{\beta} \in \Theta$ for open, convex subset $\Theta$ of $\mathcal{R}^p$)...

In this section, we provide theoretical results about the asymptotic properties of the Weighted Lasso Bootstrap. We prove its conditional consistency property and its asymptotic conditional distributions under certain conditions. We will also show that it has model selection consistency by conditioning on data. First, we introduce some notations.

**Notations:** The symbol "$\overset{d}{\approx}$" denotes "approximately distributed". For any matrix $A$, $\gamma_{\min}(A)$ refers to the smallest eigenvalue of $A$, whereas $\gamma_{\max}(A)$ refers to the largest eigenvalue of $A$.

Next, without loss of generality, suppose that out of the $p$ predictors, $q$ of them are relevant, and the remaining $p-q$ of them irrelevant. Obviously, we are only interested in the non-trivial case where $1 \leq q \leq p$. Then the columns of $X$ can be partitioned into

$$X = \begin{bmatrix} X_{(1)} & X_{(2)} \end{bmatrix}$$

which corresponds to those relevant and non-relevant predictors respectively. Similarly, $\boldsymbol{\beta}_0$ can be partitioned into

$$\boldsymbol{\beta}_0 = \begin{bmatrix} \boldsymbol{\beta}_{0(1)} \\ \boldsymbol{\beta}_{0(2)} \end{bmatrix},$$

where $\boldsymbol{\beta}_{0(1)}$ refers to the $q \times 1$ vector of non-zero true regression parameters corresponding to $X_{(1)}$, and $\boldsymbol{\beta}_{0(2)}$ is a $(p-q) \times 1$ zero vector that corresponds to $X_{(2)}$. In addition, define

$$\begin{bmatrix} C_{n(11)} & C_{n(12)} \\ C_{n(21)} & C_{n(22)} \end{bmatrix} := \frac{1}{n} X'X = \frac{1}{n} \begin{bmatrix} X'_{(1)}X_{(1)} & X'_{(1)}X_{(2)} \\ X'_{(2)}X_{(1)} & X'_{(2)}X_{(2)} \end{bmatrix}$$

which, by our assumption (4), converges to

$$C = \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix}.$$

Furthermore, let $D_n = diag(W_1, \ldots, W_n)$, and define

$$\begin{bmatrix} C^w_{n(11)} & C^w_{n(12)} \\ C^w_{n(21)} & C^w_{n(22)} \end{bmatrix} := \frac{1}{n} X' D_n X = \frac{1}{n} \begin{bmatrix} X'_{(1)} D_n X_{(1)} & X'_{(1)} D_n X_{(2)} \\ X'_{(2)} D_n X_{(1)} & X'_{(2)} D_n X_{(2)} \end{bmatrix}.$$

Finally, an estimator $\widehat{\boldsymbol{\beta}}_n$ is said to be equal in sign to the true parameter $\boldsymbol{\beta}_0$, if

$$\operatorname{sgn}(\widehat{\boldsymbol{\beta}}_n) = \operatorname{sgn}(\boldsymbol{\beta}_0),$$

and is denoted as

$$\widehat{\boldsymbol{\beta}}_n \overset{s}{=} \boldsymbol{\beta}_0.$$

An important lemma, which underlies all our asymptotic results, is given below.

**Lemma 2.1.** ***(Conditional Slutsky's)*** *Consider two sequences $\{V_n\}$ and $\{U_n\}$ and two other random variables $V$ and $U$, all defined on the same product space $(\Omega, \mathcal{F})$. If*

$$V_n | data \xrightarrow{c.p.} V \qquad and \qquad U_n | data \xrightarrow{c.d.} U,$$

*then*

$$(V_n U_n) | data \xrightarrow{c.d.} VU \qquad and \qquad (V_n + U_n) | data \xrightarrow{c.d.} V + U.$$

*Proof.* For each fixed infinite sequence of data, the results follow from properties of convergence in distribution due to Slutsky's theorem. □

**Theorem 2.1.** *Consider the linear regression model in* (2)*, with the predictor matrix $X$ satisfying* (3) *and* (4)*.*

*(a)* ***(Conditional Consistency)*** *If $\frac{\lambda_n}{n} \to 0$, then*

$$\widehat{\boldsymbol{\beta}}^w_n | data \xrightarrow{c.p.} \boldsymbol{\beta}_0.$$

*(b) If $\frac{\lambda_n}{n} \to \lambda_0 \in (0, \infty)$, then*

$$\left(\widehat{\boldsymbol{\beta}}^w_n - \boldsymbol{\beta}_0\right) \Big| data \xrightarrow{c.p.} \arg\min g,$$

*where*

$$g(\boldsymbol{u}) = \boldsymbol{u}' C \boldsymbol{u} + \lambda_0 \|\boldsymbol{\beta}_0 + \boldsymbol{u}\|_1.$$

**Theorem 2.2.** ***(Asymptotic Conditional Distribution)*** *Consider the linear regression model in* (2)*, with the predictor matrix $X$ satisfying* (3) *and* (4)*. In addition, we assume that*

$$\mathbb{E}\left(\epsilon_i^4\right) < \infty \quad \forall \quad i. \tag{8}$$

*Let $\widehat{\boldsymbol{\beta}}_n^{OLS}$ be the ordinary least squares (OLS) estimator for $\boldsymbol{\beta}$ in the linear model* (2) *with the aforementioned assumptions.*

*(a)* *If $\dfrac{\lambda_n}{\sqrt{n}} \to 0$, then*

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_n^w - \widehat{\boldsymbol{\beta}}_n^{OLS}\right)\Bigg|\, data \xrightarrow{c.d.} N\left(\mathbf{0}, \sigma^2 C^{-1}\right).$$

*(b)* *If $\dfrac{\lambda_n}{\sqrt{n}} \to \lambda_0 \in (0, \infty)$, then*

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_n^w - \widehat{\boldsymbol{\beta}}_n^{OLS}\right)\Bigg|\, data \overset{d}{\approx} \arg\min\left(V_n^* \big|\, data\right),$$

*where*

$$V_n^*(\boldsymbol{u}) = -2\boldsymbol{u}'\Psi + \boldsymbol{u}'C\boldsymbol{u} + \lambda_0 \sum_{j=1}^{p}\left[u_j\, sgn(\widehat{\beta}_{n,j})\mathbb{1}_{\{\widehat{\beta}_{n,j}\neq 0\}} + |u_j|\mathbb{1}_{\{\widehat{\beta}_{n,j}=0\}}\right],$$

*for $\Psi \sim N\left(\mathbf{0}, \sigma^2 C\right)$.*

*(c)* *Suppose we further assume that the non-invertible matrix $C$ satisfies the following assumptions:*

$$C_{12} = \mathbf{0} \qquad \text{and} \qquad \frac{\gamma_{min}(C)}{\gamma_{max}(C_{11})} > \frac{q-1}{q}.$$

*Let $\widehat{\boldsymbol{\beta}}_n^{LAS}$ be the corresponding strongly consistent Lasso estimator with its own penalty term $\lambda_n^*$ that satisfies*

$$\lambda_n^* = \mathcal{O}(n^{\delta_2}) \qquad \text{for some} \quad \frac{1}{2} < \delta_2 < 1.$$

*If* $\dfrac{\lambda_n}{\sqrt{n}} \to \lambda_0 \in (0, \infty)$, *then*

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_n^w - \widehat{\boldsymbol{\beta}}_n^{LAS}\right)\Big| data \overset{d}{\approx} \arg\min\left(V_n^{**}\big| data\right),$$

*where*

$$V_n^{**}(\boldsymbol{u}) = -2\boldsymbol{u}'\Psi_n + \boldsymbol{u}'C\boldsymbol{u} + \lambda_0 \sum_{j=1}^{p}\left[u_j\, sgn(\beta_{0,j})\mathbb{1}_{\{\beta_{0,j}\neq 0\}} + |u_j|\mathbb{1}_{\{\beta_{0,j}=0\}}\right],$$

*for*

$$\Psi_n \sim N\left(\frac{1}{\sqrt{n}}X'\boldsymbol{e}_n^{LAS}, \sigma^2 C\right),$$

*and* $\boldsymbol{e}_n^{LAS} = Y - X\widehat{\boldsymbol{\beta}}_n^{LAS}$ .

**Theorem 2.3.** ***(Conditional Model Selection Consistency)*** *Consider the linear regression model in* (2)*, with assumptions* (3)*,* (4) *and* (8)*. In addition, assume the* ***strong irrepresentable conditions*** *(Zhao and Yu, 2006): There exists a positive constant vector* $\boldsymbol{\eta}$ *such that*

$$\left|C_{n(21)}\left(C_{n(11)}\right)^{-1} sgn\left(\boldsymbol{\beta}_{0(1)}\right)\right| \leq \boldsymbol{J} - \boldsymbol{\eta}, \tag{9}$$

*where* $\boldsymbol{J}$ *is a* $(p \times q)$ *vector of ones and the inequality holds element wise. Then, for all* $\lambda_n$ *that satisfies*

$$\lambda_n = \mathcal{O}(n^c) \qquad \text{for some} \quad \frac{1}{2} < c < 1,$$

*we have*

$$P\left(\widehat{\boldsymbol{\beta}}_n^w(\lambda_n) \overset{s}{=} \boldsymbol{\beta}_0\Big| data\right) = 1 - o\left(e^{-n^{2c-1}}\right),$$

*and hence*

$$P\left(\textit{WLB selects true model}\,\big| data\right) \geq P\left(\widehat{\boldsymbol{\beta}}_n^w(\lambda_n) \overset{s}{=} \boldsymbol{\beta}_0\Big| data\right) \to 1 \;\; \textit{as} \;\; n \to \infty.$$

**Corollary 2.3.1.** *Zhao and Yu (2006) established sufficient conditions for the Strong Irrepresentable condition in Theorem 2.3. These sufficient conditions are basically additional assumptions on the predictor matrix* $X$*. We refer readers to Zhao and Yu (2006) for a proof of these sufficient conditions.*

**Discussion on Theorem 2.2:** With our assumption (4), $\widehat{\boldsymbol{\beta}}_n^{\text{OLS}}$ is a strongly consistent estimator for $\boldsymbol{\beta}$ in (2) (See, for example, Lai et al. (1978)). In fact, $\widehat{\boldsymbol{\beta}}_n^{\text{OLS}}$ in parts (a) and (b) of Theorem 2.2 can be replaced by any strongly consistent estimator $\widehat{\boldsymbol{\beta}}_n$ that satisfies

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_n - \widehat{\boldsymbol{\beta}}_n^{\text{OLS}}\right)\Bigg|\text{data} \xrightarrow{\text{c.p.}} \mathbf{0}. \tag{10}$$

Any regular Lasso estimator $\widehat{\boldsymbol{\beta}}_n^{\text{LAS}}$ with a penalty term $\lambda_n^* = o(\sqrt{n})$ could be a candidate for $\widehat{\boldsymbol{\beta}}_n$ in this case. The asterisk in $\lambda_n^*$ is to distinguish it from the penalty term $\lambda_n$ that we pick for the Weighted Lasso bootstrap. However, this rate of convergence for $\lambda_n^*$ is different from the one required in part (c) of Theorem 2.2. The additional assumptions on the matrix $C$ in part(c) of the theorem can be achieved by specific experimental design (Chatterjee and Lahiri, 2011). Part (c) of the theorem is to illustrate the fact that we could capitalize on the properties of Lasso estimators to remove the dependency of penalty term on sample path in part (b) of the theorem, but in the process, we introduce another term that depends on the sample path in the location parameter (mean) of $\Psi_n$. We refer readers to the Appendix section for a more detailed derivation and discussion of Theorem 2.2.

**Discussion on Theorem 2.3:** We establish that the Weighted Lasso Bootstrap exhibits conditional model selection consistency given data. Loosely-speaking, we would like to show that as $n$ gets larger, it becomes more likely for the zero elements in $\widehat{\boldsymbol{\beta}}_n^w$ to match those in $\boldsymbol{\beta}_0$. It is imperative to note that, in general, consistency in parameter estimation does not translate into model selection consistency, and vice versa. A typical example would be the OLS approach, where it produces consistent estimators but generally performs poorly in model selection. Our approach is similar to Zhao and Yu (2006), where we show that the Weighted Lasso Bootstrap exhibits conditional sign consistency, which then implies conditional model selection consistency.

# 3 Appendix

Here are the proofs for the theorems in this paper.

**Lemma 3.1.** *Assume* (3) *and* (4). *Then, as* $n \to \infty$,

$$\frac{1}{n} X' D_n X \xrightarrow{a.s.} C \tag{11}$$

*Proof.* Note that $\frac{1}{n} X'X$ always has a fixed $p \times p$ dimension. Coupled with assumption (3), the Strong Law of Large Numbers ensures that

$$\frac{1}{n} X'(D_n - I)X = \frac{1}{n} \sum_{i=1}^{n} (W_i - 1) \boldsymbol{x}_i \boldsymbol{x}_i' \xrightarrow{\text{a.s.}} \mathbf{0}.$$

Assumption (4) tells us that

$$\frac{1}{n} X'X \to C.$$

Therefore, by Continuous Mapping Theorem,

$$\frac{1}{n} X' D_n X = \frac{1}{n} X'(D_n - I)X + \frac{1}{n} X'X \xrightarrow{\text{a.s.}} C$$

□

**Lemma 3.2.** *Assume* (3) *and* (4). *Furthermore, let the error terms* $\{\epsilon_i\}$ *be iid with mean 0 and variance* $\sigma^2$. *Then,*

$$\left( \frac{1}{n} X' D_n \boldsymbol{\epsilon} \right) \Bigg| data \xrightarrow{c.p.} \mathbf{0}. \tag{12}$$

*Proof.* First, by Jensen's inequality,

$$\mathbb{E}|X| = \mathbb{E}(\sqrt{X^2}) \leq \sqrt{\mathbb{E}(X^2)} \leq \mathbb{E}X^2 = \sigma^2 < \infty.$$

Coupled with assumption (3), conditional on data,

$$\frac{1}{n} \sum_{i=1}^{n} \|\epsilon_i \boldsymbol{x}_i\|_2 \leq \max_{1 \leq i \leq n} \|\boldsymbol{x}_i\|_2 \times \frac{1}{n} \sum_{i=1}^{n} |\epsilon_i| < \infty \tag{13}$$

for almost every sample path $\omega$. In addition, since $\mathbb{E}(\epsilon_i) = 0$, Lemma 3.1 of Chatterjee and Lahiri (2011) gives

$$\frac{1}{n} \sum_{i=1}^{n} \epsilon_i \boldsymbol{x}_i \xrightarrow{\text{a.s.}} \mathbf{0},$$

which implies that

$$\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\boldsymbol{x}_i\bigg|\text{data}\xrightarrow{\text{c.p.}}\mathbf{0}. \tag{14}$$

Furthermore, note that by conditioning on data,

$$\begin{aligned}\frac{1}{n}\max_{1\le i\le n}\|\epsilon_i\boldsymbol{x}_i\|_2 &= \frac{1}{n}\max_{1\le i\le n}|\epsilon_i|\|\boldsymbol{x}_i\|_2\\ &\le \frac{1}{n}\max_{1\le i\le n}|\epsilon_i|\times\max_{1\le i\le n}\|\boldsymbol{x}_i\|_2\\ &\xrightarrow{\text{c.p.}} 0,\end{aligned} \tag{15}$$

where the last line follows from the fact that

$$\mathbb{E}|\epsilon_1|<\infty \implies \frac{1}{n}\max_{1\le i\le n}|\epsilon_i|\xrightarrow{\text{a.s.}}0$$

by Lemma 14 of Newton (1991).

Conditional on data, $\boldsymbol{\epsilon}$ is fixed albeit unobservable. Hence, for every $t>0$, by the multi-dimensional Chebyshev's inequality,

$$\begin{aligned}&P\left(\left\|\frac{1}{n}X'D_n\boldsymbol{\epsilon}-\mathbf{0}\right\|_2\ge t\middle|\text{data}\right)\\ &\le\frac{1}{t^2}\mathbb{E}_{\boldsymbol{W}}\left\{\left\|\frac{1}{n}X'D_n\boldsymbol{\epsilon}\right\|_2^2\middle|\text{data}\right\}\\ &=\frac{1}{t^2}\mathbb{E}_{\boldsymbol{W}}\left\{\left\|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\boldsymbol{x}_iW_i\right\|_2^2\quad\text{given data}\right\}\\ &\le\frac{2}{t^2}\left\{\mathbb{E}_{\boldsymbol{W}}\left\|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\boldsymbol{x}_i(W_i-1)\right\|_2^2+\left\|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\boldsymbol{x}_i\right\|_2^2\quad\text{given data}\right\}\\ &=\frac{2}{t^2}\left\{\frac{1}{n^2}\sum_{i=1}^{n}\|\epsilon_i\boldsymbol{x}_i\|_2^2+\left\|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\boldsymbol{x}_i\right\|_2^2\quad\text{given data}\right\}\\ &\le\frac{2}{t^2}\left\{\frac{1}{n}\max_{1\le i\le n}\|\epsilon_i\boldsymbol{x}_i\|_2\times\frac{1}{n}\sum_{i=1}^{n}\|\epsilon_i\boldsymbol{x}_i\|_2+\left\|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\boldsymbol{x}_i\right\|_2^2\quad\text{given data}\right\}\end{aligned}$$

$$\to 0$$

due to (13), (14) and (15). Finally, by Lemma 3 of Newton (1991),

$$\left(\frac{1}{n}X'D_n\boldsymbol{\epsilon}\right)\Big|\text{data} \xrightarrow{\text{c.p.}} \mathbf{0}.$$

□

Now we are ready to prove Theorem 2.1.

*Proof of Theorem 2.1.* From (7), conditional on data,

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_n^w &= \arg\min_{\boldsymbol{\beta}} \frac{1}{\overline{W}}\left\{\frac{1}{n}(Y - X\boldsymbol{\beta})'D_n(Y - X\boldsymbol{\beta}) + \frac{\lambda_n}{n}\|\boldsymbol{\beta}\|_1\right\} \\ &= \arg\min_{\boldsymbol{\beta}} \frac{1}{\overline{W}}\left\{\frac{1}{n}[\boldsymbol{\epsilon} - X(\boldsymbol{\beta} - \boldsymbol{\beta}_0)]'D_n[\boldsymbol{\epsilon} - X(\boldsymbol{\beta} - \boldsymbol{\beta}_0)] + \frac{\lambda_n}{n}\|\boldsymbol{\beta}_0 + \boldsymbol{\beta} - \boldsymbol{\beta}_0\|_1\right\}.\end{aligned}$$

Therefore,

$$\begin{aligned}(\widehat{\boldsymbol{\beta}}_n^w - \boldsymbol{\beta}_0) &= \arg\min_{\boldsymbol{u}} \frac{1}{\overline{W}}\left\{\frac{1}{n}(\boldsymbol{\epsilon} - X\boldsymbol{u})'D_n(\boldsymbol{\epsilon} - X\boldsymbol{u}) + \frac{\lambda_n}{n}\|\boldsymbol{\beta}_0 + \boldsymbol{u}\|_1\right\} \\ &= \arg\min_{\boldsymbol{u}} \frac{1}{\overline{W}}\left\{\frac{1}{n}[-2\boldsymbol{u}'(X'D_n\boldsymbol{\epsilon}) + \boldsymbol{u}'(X'D_nX)\boldsymbol{u}] + \frac{\lambda_n}{n}\|\boldsymbol{\beta}_0 + \boldsymbol{u}\|_1\right\}.\end{aligned}$$

The strong law of large numbers ensures that $\overline{W} \xrightarrow{\text{a.s.}} 1$. Let

$$g_n(\boldsymbol{u}) := -2\boldsymbol{u}'\left(\frac{X'D_n\boldsymbol{\epsilon}}{n}\right) + \boldsymbol{u}'\left(\frac{X'D_nX}{n}\right)\boldsymbol{u} + \frac{\lambda_n}{n}\|\boldsymbol{\beta}_0 + \boldsymbol{u}\|_1.$$

By Lemma 3.1, we have

$$\frac{1}{n}X'D_nX \xrightarrow{\text{p}} C.$$

By Lemma 3.2, we have

$$\left(\frac{1}{n}X'D_n\boldsymbol{\epsilon}\right)\Big|\text{data} \xrightarrow{\text{c.p.}} \mathbf{0}.$$

Hence, if $\frac{\lambda_n}{n} \to \lambda_0 \in [0, \infty)$, then by Lemma 2.1,

$$g_n(\boldsymbol{u})\big|\text{data} \xrightarrow{\text{c.p.}} g(\boldsymbol{u}) \equiv \boldsymbol{u}'C\boldsymbol{u} + \lambda_0\|\boldsymbol{\beta}_0 + \boldsymbol{u}\|.$$

Note that $g_n(\boldsymbol{u})$ is a sequence of random convex functions of $\boldsymbol{u}$. Hence, by the Convexity Lemma (Pollard, 1991), for a compact set $K \subset \Theta$, where $\Theta$ is itself a convex, open subset of $\mathcal{R}^p$,

$$\sup_{\boldsymbol{u} \in K \subset \Theta} |g_n(\boldsymbol{u}) - g(\boldsymbol{u})| \Big| \text{data} \xrightarrow{\text{c.p.}} 0.$$

Also, note that $\left( \widehat{\boldsymbol{\beta}}_n^w \big| \text{data} \right) = O_p(1)$. Therefore,

$$\begin{aligned}
&\left( \widehat{\boldsymbol{\beta}}_n^w - \boldsymbol{\beta}_0 \right) \Bigg| \text{data} \\
&= \arg\min_{\boldsymbol{u}} \left\{ \frac{1}{\overline{W}} g_n(\boldsymbol{u}) \Big| \text{data} \right\} \\
&\xrightarrow{\text{c.p.}} \arg\min_{\boldsymbol{u}} g(\boldsymbol{u}).
\end{aligned}$$

It follows that if $\lambda_0 = 0$, then $\arg\min_{\boldsymbol{u}} g(\boldsymbol{u}) = \boldsymbol{0}$, i.e. $\widehat{\boldsymbol{\beta}}_n^w \big| \text{data} \xrightarrow{\text{c.p.}} \boldsymbol{\beta}_0$. □

**Lemma 3.3.** *Assume* (3) *and* (4). *Let* $\widehat{\boldsymbol{\beta}}_n$ *be a strongly consistent estimator for* $\boldsymbol{\beta}$ *in the linear model* (2) *with assumptions* (3) *and* (4). *Then*

$$\left( \frac{1}{n} \sum_{i=1}^{n} e_i^2 \boldsymbol{x}_i \boldsymbol{x}_i' \right) \Bigg| data \xrightarrow{c.p.} \sigma^2 C,$$

*where* $e_i = Y_i - \boldsymbol{x}_i' \widehat{\boldsymbol{\beta}}_n$.

*Proof.* Without loss of generality, we first consider the univariate case.

$$\begin{aligned}
e_i^2 &= \left( Y_i - \widehat{Y}_i \right)^2 \\
&= Y_i^2 + x_i^2 \left( \widehat{\beta}_n \right)^2 - 2\widehat{\beta}_n x_i Y_i \\
&= (x_i \beta_0 + \epsilon_i)^2 + x_i^2 \left( \widehat{\beta}_n \right)^2 - 2\widehat{\beta}_n x_i (x_i \beta_0 + \epsilon_i) \\
&= \epsilon_i^2 + x_i^2 \left[ \beta_0^2 + \left( \widehat{\beta}_n \right)^2 - 2 \left( \beta_0 \widehat{\beta}_n \right) \right] - 2 x_i \epsilon_i \left( \beta_0 \widehat{\beta}_n \right),
\end{aligned}$$

which leads us to

$$\frac{1}{n} \sum_{i=1}^{n} x_i^2 e_i^2$$

$$= \frac{1}{n}\sum_{i=1}^{n} x_i^2\epsilon_i^2 + \left[\beta_0^2 + \left(\widehat{\beta}_n\right)^2 - 2\left(\beta_0\widehat{\beta}_n\right)\right]\left(\frac{1}{n}\sum_{i=1}^{n} x_i^4\right) - 2\left(\beta_0\widehat{\beta}_n\right)\left(\frac{1}{n}\sum_{i=1}^{n} x_i^3\epsilon_i\right).$$

Continuous Mapping Theorem ensures that

$$\beta_0^2 + \left(\widehat{\beta}_n\right)^2 - 2\left(\beta_0\widehat{\beta}_n\right) \xrightarrow{\text{a.s.}} 0,$$

and

$$\beta_0\widehat{\beta}_n \xrightarrow{\text{a.s.}} \beta_0^2.$$

By assumption (3), we have

$$\frac{1}{n}\sum_{i=1}^{n} x_i^4 = \mathcal{O}(1).$$

By assumption (3) and the Strong Law of Large Numbers, we have

$$\frac{1}{n}\sum_{i=1}^{n} x_i^3\epsilon_i \xrightarrow{\text{a.s.}} 0 \qquad \text{and} \qquad \frac{1}{n}\sum_{i=1}^{n} x_i^2(\epsilon_i - \sigma^2) \xrightarrow{\text{a.s.}} 0.$$

Meanwhile, by assumption (4), we have

$$\frac{1}{n}\sum_{i=1}^{n} x_i^2 \to c \quad \text{for some } c > 0.$$

Therefore, by Continuous Mapping Theorem,

$$\frac{1}{n}\sum_{i=1}^{n} x_i^2\epsilon_i^2 \xrightarrow{\text{a.s.}} \sigma^2 c.$$

Finally, piecing the terms together, we have

$$\frac{1}{n}\sum_{i=1}^{n} x_i^2 e_i^2 \xrightarrow{\text{a.s.}} \sigma^2 c,$$

and hence

$$\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2 e_i^2\right)\Bigg|\text{data} \xrightarrow{\text{c.p.}} \sigma^2 c.$$

A sketch of proof is also provided for the multivariate case.

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} e_i^2 \boldsymbol{x}_i \boldsymbol{x}_i' \\
&= \frac{1}{n}\sum_{i=1}^{n} \epsilon_i^2 \boldsymbol{x}_i \boldsymbol{x}_i' \\
&+ \frac{2}{n}\sum_{i=1}^{n} \epsilon_i \boldsymbol{x}_i' \left(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\right) \boldsymbol{x}_i \boldsymbol{x}_i' \\
&+ \frac{1}{n}\sum_{i=1}^{n} \left\{ \boldsymbol{x}_i' \left[\boldsymbol{\beta}_0 \boldsymbol{\beta}_0' - 2\widehat{\boldsymbol{\beta}}_n \boldsymbol{\beta}_0' + \widehat{\boldsymbol{\beta}}_n \left(\widehat{\boldsymbol{\beta}}_n\right)'\right] \boldsymbol{x}_i \right\} \boldsymbol{x}_i \boldsymbol{x}_i'.
\end{aligned}$$

With our assumptions, the Strong Law of Large Numbers ensures that the first term converges to $\sigma^2 C$ with probability 1 whereas the other two terms converges to zero matrix almost surely. Therefore,

$$\left(\frac{1}{n}\sum_{i=1}^{n} e_i^2 \boldsymbol{x}_i \boldsymbol{x}_i'\right)\Bigg|\text{data} \xrightarrow{\text{c.p.}} \sigma^2 C.$$

□

**Lemma 3.4.** *Assume* (3), (4), *and* (8). *Let $\widehat{\boldsymbol{\beta}}_n^{OLS}$ be the OLS estimator for $\boldsymbol{\beta}$ in the linear model* (2) *with assumptions* (3) *and* (4). *Then,*

$$\left(\frac{1}{\sqrt{n}} X' D_n \boldsymbol{e}_n^{OLS}\right)\Bigg| data \xrightarrow{c.d.} N\left(\mathbf{0}, \sigma^2 C\right).$$

*Proof.* First, note that $\widehat{\boldsymbol{\beta}}_n^{\text{OLS}}$ is a strongly consistent estimator under assumption (4) (Lai, Robbins, and Wei, 1978). Next,

$$\begin{aligned}
&\frac{1}{\sqrt{n}} X' D_n \boldsymbol{e}_n^{\text{OLS}} \\
&= \left(\frac{1}{n}\sum_{i=1}^{n} e_i^2 \boldsymbol{x}_i \boldsymbol{x}_i'\right)^{1/2} \times \left(\sum_{i=1}^{n} e_i^2 \boldsymbol{x}_i \boldsymbol{x}_i'\right)^{-1/2} \times \sum_{i=1}^{n} e_i \boldsymbol{x}_i W_i \\
&= \left(\frac{1}{n}\sum_{i=1}^{n} e_i^2 \boldsymbol{x}_i \boldsymbol{x}_i'\right)^{1/2} \times \left(\sum_{i=1}^{n} e_i^2 \boldsymbol{x}_i \boldsymbol{x}_i'\right)^{-1/2} \times \sum_{i=1}^{n} e_i \boldsymbol{x}_i (W_i - 1),
\end{aligned}$$

where the last equality follows from the fact that

$$\begin{aligned}\sum_{i=1}^{n} e_i \boldsymbol{x}_i &= X' \boldsymbol{e}_n^{\text{OLS}} \\ &= X'Y - X'X(X'X)^{-1}X'Y \\ &= \mathbf{0}.\end{aligned}$$

By Lemma 3.3,

$$\left(\frac{1}{n}\sum_{i=1}^{n} e_i^2 \boldsymbol{x}_i \boldsymbol{x}_i'\right)^{1/2} \Bigg| \text{data} \xrightarrow{\text{c.p.}} \sigma C^{1/2}.$$

Without loss of generality, we will continue our proof for the univariate case. We shall show that the Lindeberg's Central Limit Theorem gives

$$\left(\frac{\sum_{i=1}^{n} e_i x_i (W_i - 1)}{\sqrt{\sum_{i=1}^{n} e_i^2 x_i^2}}\right) \Bigg| \text{data} \xrightarrow{\text{c.d.}} N(0,1)$$

by verifying the following Liapounov's sufficient condition

$$\frac{\sum_{i=1}^{n} \mathbb{E}\left[e_i^4 x_i^4 (W_i - 1)^4 | \text{data}\right]}{\left(Var\left[\sum_{i=1}^{n} e_i x_i (W_i - 1) | \text{data}\right]\right)^2} \to 0 \quad \text{as } n \to \infty.$$

With assumptions (3) and (8), we can use similar reasoning in Lemma 3.3 to show that

$$\sum_{i=1}^{n} e_i^4 x_i^4 = \mathcal{O}(n) \text{ a.s.},$$

since $\sum_{i=1}^{n} e_i^4 x_i^4$ can be expanded into

$$\sum_{i=1}^{n}\left[x_i^4 \epsilon_i^4 - 4x_i^5(\widehat{\beta}_n - \beta_0)\epsilon_i^3 + 6x_i^6(\widehat{\beta}_n - \beta_0)^2\epsilon_i^2 - 4x_i^7(\widehat{\beta}_n - \beta_0)^3\epsilon_i + x_i^8(\widehat{\beta}_n - \beta_0)^4\right].$$

Since $\mathbb{E}\left[(W_i - 1)^4\right] = 9$,

$$\sum_{i=1}^{n} \mathbb{E}\left[e_i^4 x_i^4 (W_i - 1)^4 | \text{data}\right] = \mathcal{O}(n)$$

On the other hand, by Lemma 3.3 and Continuous Mapping Theorem,

$$\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2 e_i^2\right)^2 \xrightarrow{\text{a.s.}} \sigma^4 c^2,$$

which implies that

$$\left(\sum_{i=1}^{n} x_i^2 e_i^2\right)^2 = \mathcal{O}(n^2) \text{ a.s.}$$

Hence,

$$\begin{aligned}
&\left(Var\left[\sum_{i=1}^{n} e_i x_i (W_i - 1)\middle|\text{data}\right]\right)^2 \\
&= \left(\sum_{i=1}^{n} x_i^2 e_i^2\right)^2 \middle|\text{data} \\
&= \mathcal{O}(n^2)
\end{aligned}$$

Therefore, conditional on data,

$$\sum_{i=1}^{n} e_i^4 x_i^4 \mathbb{E}\left[(W_i - 1)^4\right] = o\left[\left(\sum_{i=1}^{n} e_i^2 x_i^2\right)^2\right],$$

thus satisfying the Liapounov's sufficient condition. Finally, we apply Lemma 2.1 to obtain

$$\left(\frac{1}{\sqrt{n}} X' D_n \boldsymbol{e}_n^{\text{OLS}}\right)\middle|\text{data} \xrightarrow{\text{c.d.}} N\left(\mathbf{0}, \sigma^2 C\right).$$

□

**Lemma 3.5.** *Assume* (3)*,* (4)*, and* (8)*.*

*(a) Let $\widehat{\boldsymbol{\beta}}_n \neq \widehat{\boldsymbol{\beta}}_n^{OLS}$ be any other strongly consistent estimator for $\boldsymbol{\beta}$ in the linear model* (2) *under the aforementioned assumptions, and satisfies* (10)*. Then,*

$$\left(\frac{1}{\sqrt{n}} X' D_n \boldsymbol{e}_n\right)\middle| data \xrightarrow{c.d.} N\left(\mathbf{0}, \sigma^2 C\right),$$

*where $\boldsymbol{e}_n = Y - X\widehat{\boldsymbol{\beta}}_n$.*

*(b) A Lasso estimator $\widehat{\boldsymbol{\beta}}_n^{LAS}$ with a penalty term $\lambda_n^* = o(\sqrt{n})$ satisfies* (10)*, and hence can be a substitute for $\widehat{\boldsymbol{\beta}}_n^{OLS}$ in parts (a) and (b) of Theorem 2.2.*

*Proof.* First, note that

$$\left(\frac{1}{\sqrt{n}}X'(D_n - I)\boldsymbol{e}_n\right)\Big|\text{data} \xrightarrow{\text{c.d.}} N\left(\mathbf{0}, \sigma^2 C\right).$$

by Lemmas 3.3 and 3.4. Now

$$\frac{1}{\sqrt{n}}X'D_n\boldsymbol{e}_n = \frac{1}{\sqrt{n}}X'(D_n - I)\boldsymbol{e}_n + \frac{1}{\sqrt{n}}X'\boldsymbol{e}_n,$$

where, by conditioning on data,

$$\begin{aligned}\frac{1}{\sqrt{n}}X'\boldsymbol{e}_n &= \frac{1}{\sqrt{n}}X'\boldsymbol{e}_n^{\text{OLS}} + \frac{1}{\sqrt{n}}X'\left[\boldsymbol{e}_n - \boldsymbol{e}_n^{\text{OLS}}\right]\\ &= \frac{1}{\sqrt{n}}X'X\left[\widehat{\boldsymbol{\beta}}_n^{\text{OLS}} - \widehat{\boldsymbol{\beta}}_n\right]\\ &= \frac{1}{n}X'X \times \sqrt{n}\left[\widehat{\boldsymbol{\beta}}_n^{\text{OLS}} - \widehat{\boldsymbol{\beta}}_n\right]\\ &\xrightarrow{\text{c.p.}} \mathbf{0},\end{aligned}$$

by our assumption in part (a) of the Lemma. Finally, by Lemma 2.1,

$$\left(\frac{1}{\sqrt{n}}X'D_n\boldsymbol{e}_n\right)\Big|\text{data} \xrightarrow{\text{c.d.}} N\left(\mathbf{0}, \sigma^2 C\right).$$

For part (b) of the Lemma, first define

$$Q_n^{\text{LAS}}(\boldsymbol{z}) := \|(\boldsymbol{y} - X\boldsymbol{z})\|_2^2 + \lambda_n^*\|\boldsymbol{z}\|_1,$$

which leads to

$$\begin{aligned}Q_n^{\text{LAS}}\left(\widehat{\boldsymbol{\beta}}_n^{\text{OLS}} + \frac{1}{\sqrt{n}}\boldsymbol{u}\right) &= \left\|Y - X\left(\widehat{\boldsymbol{\beta}}_n^{\text{OLS}} + \frac{1}{\sqrt{n}}\boldsymbol{u}\right)\right\|_2^2 + \lambda_n^*\left\|\widehat{\boldsymbol{\beta}}_n^{\text{OLS}} + \frac{1}{\sqrt{n}}\boldsymbol{u}\right\|_1\\ &= \left\|\boldsymbol{e}_n^{\text{OLS}} - \frac{1}{\sqrt{n}}X\boldsymbol{u}\right\|_2^2 + \lambda_n^*\left\|\widehat{\boldsymbol{\beta}}_n^{\text{OLS}} + \frac{1}{\sqrt{n}}\boldsymbol{u}\right\|_1,\end{aligned}$$

and

$$\begin{aligned}Q_n^{\text{LAS}}\left(\widehat{\boldsymbol{\beta}}_n^{\text{OLS}}\right) &= \left\|Y - X\widehat{\boldsymbol{\beta}}_n^{\text{OLS}}\right\|_2^2 + \lambda_n^*\left\|\widehat{\boldsymbol{\beta}}_n^{\text{OLS}}\right\|_1\\ &= \|\boldsymbol{e}_n^{\text{OLS}}\|_2^2 + \lambda_n^*\left\|\widehat{\boldsymbol{\beta}}_n^{\text{OLS}}\right\|_1.\end{aligned}$$

Now define

$$h_n(\boldsymbol{u}) := Q_n^{\text{LAS}}\left(\widehat{\boldsymbol{\beta}}_n^{\text{OLS}} + \frac{1}{\sqrt{n}}\boldsymbol{u}\right) - Q_n^{\text{LAS}}\left(\widehat{\boldsymbol{\beta}}_n^{\text{OLS}}\right),$$

such that

$$\arg\min_{\boldsymbol{u}} h_n(\boldsymbol{u}) = \arg\min_{\boldsymbol{u}} Q_n^{\text{LAS}}\left(\widehat{\boldsymbol{\beta}}_n^{\text{OLS}} + \frac{1}{\sqrt{n}}\boldsymbol{u}\right) = \sqrt{n}\left(\widehat{\boldsymbol{\beta}}_n^{\text{LAS}} - \widehat{\boldsymbol{\beta}}_n^{\text{OLS}}\right).$$

Notice that $h_n(\boldsymbol{u})$ can be simplified into

$$\boldsymbol{u}'\left(\frac{X'X}{n}\right)\boldsymbol{u} + \frac{\lambda_n^*}{\sqrt{n}}\left\{\left\|\sqrt{n}\widehat{\boldsymbol{\beta}}_n^{\text{OLS}} + \boldsymbol{u}\right\|_1 - \left\|\sqrt{n}\widehat{\boldsymbol{\beta}}_n^{\text{OLS}}\right\|_1\right\},$$

and $h_n(\boldsymbol{u}) \to h(\boldsymbol{u}) \equiv \boldsymbol{u}'C\boldsymbol{u}$ if $\lambda_n^* = o(\sqrt{n})$. Therefore, by the Convexity Lemma (Pollard, 1991), and the fact that $\widehat{\boldsymbol{\beta}}_n^{\text{LAS}} = O_p(1)$,

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_n^{\text{LAS}} - \widehat{\boldsymbol{\beta}}_n^{\text{OLS}}\right)\Big|\text{data} \xrightarrow{\text{c.p.}} \arg\min_{\boldsymbol{u}} h(\boldsymbol{u}) = \boldsymbol{0}.$$

□

Now we are ready to prove Theorem 2.2.

*Proof of Theorem 2.2.* Define

$$Q_n(\boldsymbol{z}) := \left\|D_n^{\frac{1}{2}}(\boldsymbol{y} - X\boldsymbol{z})\right\|_2^2 + \lambda_n\|\boldsymbol{z}\|_1,$$

which leads to

$$\begin{aligned} Q_n\left(\widehat{\boldsymbol{\beta}}_n^{\text{OLS}} + \frac{1}{\sqrt{n}}\boldsymbol{u}\right) &= \left\|D_n^{\frac{1}{2}}\left[Y - X\left(\widehat{\boldsymbol{\beta}}_n^{\text{OLS}} + \frac{1}{\sqrt{n}}\boldsymbol{u}\right)\right]\right\|_2^2 + \lambda_n\left\|\widehat{\boldsymbol{\beta}}_n^{\text{OLS}} + \frac{1}{\sqrt{n}}\boldsymbol{u}\right\|_1 \\ &= \left\|D_n^{\frac{1}{2}}\left(\boldsymbol{e}_n^{\text{OLS}} - \frac{1}{\sqrt{n}}X\boldsymbol{u}\right)\right\|_2^2 + \lambda_n\left\|\widehat{\boldsymbol{\beta}}_n^{\text{OLS}} + \frac{1}{\sqrt{n}}\boldsymbol{u}\right\|_1, \end{aligned}$$

and

$$Q_n\left(\widehat{\boldsymbol{\beta}}_n^{\text{OLS}}\right) = \left\|D_n^{\frac{1}{2}}\left(Y - X\widehat{\boldsymbol{\beta}}_n^{\text{OLS}}\right)\right\|_2^2 + \lambda_n\left\|\widehat{\boldsymbol{\beta}}_n^{\text{OLS}}\right\|_1$$

$$= \left\| D_n^{\frac{1}{2}} \boldsymbol{e}_n^{\text{OLS}} \right\|_2^2 + \lambda_n \left\| \widehat{\boldsymbol{\beta}}_n^{\text{OLS}} \right\|_1 .$$

Now define

$$V_n(\boldsymbol{u}) := Q_n \left( \widehat{\boldsymbol{\beta}}_n^{\text{OLS}} + \frac{1}{\sqrt{n}} \boldsymbol{u} \right) - Q_n \left( \widehat{\boldsymbol{\beta}}_n^{\text{OLS}} \right),$$

such that

$$\arg\min_{\boldsymbol{u}} V_n(\boldsymbol{u}) = \arg\min_{\boldsymbol{u}} Q_n \left( \widehat{\boldsymbol{\beta}}_n^{\text{OLS}} + \frac{1}{\sqrt{n}} \boldsymbol{u} \right) = \sqrt{n} \left( \widehat{\boldsymbol{\beta}}_n^{w} - \widehat{\boldsymbol{\beta}}_n^{\text{OLS}} \right).$$

Notice that $V_n(\boldsymbol{u})$ can be simplified into

$$-2\boldsymbol{u}' \left( \frac{X' D_n \boldsymbol{e}_n^{\text{OLS}}}{\sqrt{n}} \right) + \boldsymbol{u}' \left( \frac{X' D_n X}{n} \right) \boldsymbol{u} + \lambda_n \left\{ \left\| \widehat{\boldsymbol{\beta}}_n^{\text{OLS}} + \frac{1}{\sqrt{n}} \boldsymbol{u} \right\|_1 - \left\| \widehat{\boldsymbol{\beta}}_n^{\text{OLS}} \right\|_1 \right\}.$$

By Lemma 3.1,

$$\frac{1}{n} X' D_n X \xrightarrow{\text{p}} C.$$

By Lemma 3.4,

$$\frac{1}{\sqrt{n}} X' D_n \boldsymbol{e}_n^{\text{OLS}} \xrightarrow{\text{c.d.}} N\left(\mathbf{0}, \sigma^2 C\right).$$

For the penalty term,

$$\begin{aligned}
&\lambda_n \left\{ \left\| \widehat{\boldsymbol{\beta}}_n^{\text{OLS}} + \frac{1}{\sqrt{n}} \boldsymbol{u} \right\|_1 - \left\| \widehat{\boldsymbol{\beta}}_n^{\text{OLS}} \right\|_1 \right\} \\
&= \frac{\lambda_n}{\sqrt{n}} \sum_{j=1}^{p} \left\{ \left| \sqrt{n} \widehat{\beta}_{n,j}^{\text{OLS}} + \mu_j \right| - \left| \sqrt{n} \widehat{\beta}_{n,j}^{\text{OLS}} \right| \right\} \\
&= \frac{\lambda_n}{\sqrt{n}} \sum_{j=1}^{p} \left\{ \left| \sqrt{n} \left[ \beta_{o,j} + \left( \widehat{\beta}_{n,j}^{\text{OLS}} - \beta_{o,j} \right) \right] + \mu_j \right| - \left| \sqrt{n} \left[ \beta_{o,j} + \left( \widehat{\beta}_{n,j}^{\text{OLS}} - \beta_{o,j} \right) \right] \right| \right\} \\
&:= \frac{\lambda_n}{\sqrt{n}} \sum_{j=1}^{p} p_n(u_j).
\end{aligned}$$

First consider the case when

$$\frac{\lambda_n}{\sqrt{n}} \to \lambda_0 \in [0, \infty).$$

When $\beta_{o,j} \neq 0$, for large $n$, $\widehat{\beta}_{n,j}^{\text{OLS}} - \beta_{o,j} \xrightarrow{\text{a.s.}} 0$, and $\sqrt{n}\beta_{o,j}$ dominates $u_j$. Hence, it is easy to verify that for $\beta_{o,j} \neq 0$, $p_n(u_j)$ becomes

$$u_j \text{sgn}\left(\beta_{o,j}\right) \mathbb{1}_{\{\beta_{o,j} \neq 0\}}$$

based on the following observations

- when $\beta_{o,j} > 0$ and $u_j > 0$, then $p_n(u_j) = u_j$;
- when $\beta_{o,j} > 0$ and $u_j < 0$, then $p_n(u_j) = u_j$;
- when $\beta_{o,j} < 0$ and $u_j > 0$, then $p_n(u_j) = -u_j$;
- when $\beta_{o,j} < 0$ and $u_j < 0$, then $p_n(u_j) = -u_j$.

On the other hand,$\beta_{o,j} = 0$, $p_n(u_j)$ is back to

$$\left|\sqrt{n}\widehat{\beta}_{n,j}^{\text{OLS}} + \mu_j\right| - \left|\sqrt{n}\widehat{\beta}_{n,j}^{\text{OLS}}\right|,$$

where $\sqrt{n}\widehat{\beta}_{n,j}^{\text{OLS}}$ is normally distributed, and there is no $N > 0$ such that $n \geq N \implies \widehat{\beta}_{n,j}^{\text{OLS}} = 0$. This causes $p_n(u_j)$ to depend on sample path when $\beta_{o,j} = 0$:

$$\mathbb{1}_{\{\beta_{o,j} \neq 0\}} \left[u_j \text{sgn}\left(\widehat{\beta}_{n,j}^{\text{OLS}}\right) \mathbb{1}_{\left\{\widehat{\beta}_{n,j}^{\text{OLS}} \neq 0\right\}} + |u_j| \mathbb{1}_{\left\{\widehat{\beta}_{n,j}^{\text{OLS}} = 0\right\}}\right].$$

Hence, $p_n(u_j)$ becomes

$$u_j \text{sgn}\left(\beta_{o,j}\right) \mathbb{1}_{\{\beta_{o,j} \neq 0\}} + \mathbb{1}_{\{\beta_{o,j} \neq 0\}} \left[u_j \text{sgn}\left(\widehat{\beta}_{n,j}^{\text{OLS}}\right) \mathbb{1}_{\left\{\widehat{\beta}_{n,j}^{\text{OLS}} \neq 0\right\}} + |u_j| \mathbb{1}_{\left\{\widehat{\beta}_{n,j}^{\text{OLS}} = 0\right\}}\right].$$

Since $p_n(u_j)$ still depends on sample path anyway, we left the expression as

$$\left[u_j \, \text{sgn}(\widehat{\beta}_{n,j}) \mathbb{1}_{\{\widehat{\beta}_{n,j} \neq 0\}} + |u_j| \mathbb{1}_{\{\widehat{\beta}_{n,j} = 0\}}\right].$$

Then, by Lemma 2.1, conditional on data,

$$V_n(\boldsymbol{u}) \overset{d}{\approx} V_n^*(\boldsymbol{u}) \equiv -2\boldsymbol{u}'\Psi + \boldsymbol{u}'C\boldsymbol{u} + \lambda_0 \sum_{j=1}^{p} \left[u_j \, \text{sgn}(\widehat{\beta}_{n,j}) \mathbb{1}_{\{\widehat{\beta}_{n,j} \neq 0\}} + |u_j| \mathbb{1}_{\{\widehat{\beta}_{n,j} = 0\}}\right],$$

where $\Psi \sim N\left(\mathbf{0}, \sigma^2 C\right)$. Finally, conditional on data, $V_n(\boldsymbol{u})$ is convex, and $V_n^*(\boldsymbol{u})$ has unique minimum. Therefore, it follows from Geyer (1996) that

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_n^w - \boldsymbol{\beta}_0\right)\Big|\text{data} = \arg\min_{\boldsymbol{u}}\left\{V_n(\boldsymbol{u})\middle|\text{data}\right\} \overset{d}{\approx} \arg\min_{\boldsymbol{u}}\left\{V_n^*(\boldsymbol{u})\middle|\text{data}\right\}.$$

Notice that we will arrive at the same result for $p_n(u_j)$ even if we substitute $\widehat{\boldsymbol{\beta}}_n^{\text{OLS}}$ for a strongly consistent Lasso estimator $\widehat{\boldsymbol{\beta}}_n^{\text{LAS}}$ since Knight and Fu (2000) proved that $\sqrt{n}\left(\widehat{\beta}_{n,j}^{\text{LAS}} - \beta_{0,j}\right)$ is normally distributed, so it is not possible to obtain

$$\left\|\widehat{\boldsymbol{\beta}}_n^{\text{LAS}} - \boldsymbol{\beta}_0\right\| = \mathcal{O}(\sqrt{n}),$$

unless additional assumptions about the matrix $C$ are assumed in part (c), so that $\exists N_\omega$ such that when $\beta_{0,j} = 0$,

$$n \geq N_\omega \implies \widehat{\beta}_{n,j}^{\text{LAS}} = 0.$$

However, the penalty $\lambda_n^*$ required in part (c) has to be

$$\lambda_n^* = \mathcal{O}(n^{\delta_2}) \qquad \text{for some} \quad \frac{1}{2} < \delta_2 < 1,$$

which fails to satisfy (10). Thus, $\frac{1}{\sqrt{n}}X'\boldsymbol{e}_n^{\text{LAS}}$ may not necessarily shrink to zero, and becomes dependent on the sample path $\omega$. This contributes to a moving mean of $\Psi_n$.

Finally, if $\dfrac{\lambda_n}{\sqrt{n}} \to 0$, then

$$\begin{aligned}\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_n^w - \boldsymbol{\beta}_0\right)\Big|\text{data} &\xrightarrow{\text{c.d.}} \arg\min_{\boldsymbol{u}} V(\boldsymbol{u}) \\ &= \arg\min_{\boldsymbol{u}}\left\{-2\boldsymbol{u}'\Psi + \boldsymbol{u}'C\boldsymbol{u}\right\} \\ &= C^{-1}\Psi \sim N(\mathbf{0}, \sigma^2 C^{-1}).\end{aligned}$$

□

**Lemma 3.6.** *Assume all conditions stated in Theorem 2.3. Then*

$$P\left(\widehat{\boldsymbol{\beta}}_n^w(\lambda_n) \overset{s}{=} \boldsymbol{\beta}_0\middle|\,data\right) \geq P\left(A_n^w \cap B_n^w\middle|\,data\right),$$

*where*

$$A_n^w \equiv \left\{ \left| \left(C_{n(11)}^w\right)^{-1} \boldsymbol{Z}_{n(1)}^w \right| < \sqrt{n} \left[ \left|\boldsymbol{\beta}_{0(1)}\right| - \frac{\lambda_n}{2n} \left| \left(C_{n(11)}^w\right)^{-1} sgn\left(\boldsymbol{\beta}_{0(1)}\right) \right| \right] \textit{ element-wise} \right\}$$

$$B_n^w \equiv \left\{ \left| C_{n(21)}^w \left(C_{n(11)}^w\right)^{-1} \boldsymbol{Z}_{n(1)}^w - \boldsymbol{Z}_{n(2)}^w \right| \leq \frac{\lambda_n}{2\sqrt{n}} \left(\boldsymbol{\eta} - |\boldsymbol{\rho}_n^w|\right) \textit{ element-wise} \right\},$$

*for*

$$\begin{aligned} \boldsymbol{Z}_{n(1)}^w &= \frac{1}{\sqrt{n}} X'_{n(1)} D_n \boldsymbol{\epsilon}_n, \\ \boldsymbol{Z}_{n(2)}^w &= \frac{1}{\sqrt{n}} X'_{n(2)} D_n \boldsymbol{\epsilon}_n, \\ \boldsymbol{\rho}_n^w &= \left[ C_{n(21)}^w \left(C_{n(11)}^w\right)^{-1} - C_{n(21)} \left(C_{n(11)}\right)^{-1} \right] sgn\left(\boldsymbol{\beta}_{0(1)}\right). \end{aligned}$$

*Proof.* Based on

$$Q_n(\boldsymbol{z}) = \left\| D_n^{\frac{1}{2}} \left(Y - X\boldsymbol{z}\right) \right\|_2^2 + \lambda_n \left\|\boldsymbol{z}\right\|_1,$$

we consider $Q_n(\boldsymbol{\beta}_0 + \boldsymbol{u}_n)$ and drop the terms that do not involve $\boldsymbol{u}_n$ to obtain

$$V_n(\boldsymbol{u}_n) = -2\boldsymbol{u}_n' \left(X' D_n \boldsymbol{\epsilon}_n\right) + \boldsymbol{u}_n' \left(X' D_n X\right) \boldsymbol{u} + \lambda_n \left\{ \left\|\boldsymbol{\beta}_0 + \boldsymbol{u}_n\right\|_1 \right\},$$

such that

$$\begin{aligned} \left(\widehat{\boldsymbol{\beta}}_n^w - \boldsymbol{\beta}_0\right) \Big| \text{data} &= \arg\min_{\boldsymbol{u}_n} \left\{ Q_n(\boldsymbol{\beta}_0 + \boldsymbol{u}_n) \middle| \text{data} \right\} \\ &= \arg\min_{\boldsymbol{u}_n} \left\{ V_n(\boldsymbol{u}_n) \middle| \text{data} \right\}. \end{aligned}$$

Differentiating the first two terms with respect to $\boldsymbol{u}_n$ yields

$$\begin{aligned} & 2X' D_n X \boldsymbol{u}_n - 2X' D_n \boldsymbol{\epsilon}_n \\ &= 2\sqrt{n} \left( \frac{1}{n} X' D_n X \right) \left( \sqrt{n} \boldsymbol{u}_n \right) - 2\sqrt{n} \left( \frac{1}{\sqrt{n}} X' D_n \boldsymbol{\epsilon}_n \right) \\ &= 2\sqrt{n} \left[ C_n^w \left( \sqrt{n} \boldsymbol{u}_n \right) - \boldsymbol{Z}_n^w \right] \end{aligned}$$

Note that $\widehat{\boldsymbol{\beta}}_n^w = \widehat{\boldsymbol{u}}_n + \boldsymbol{\beta}_0$, which can be partitioned into

$$\widehat{\boldsymbol{\beta}}_n^w = \begin{bmatrix} \widehat{\boldsymbol{\beta}}_{n(1*)}^w \\ \widehat{\boldsymbol{\beta}}_{n(2*)}^w \end{bmatrix},$$

where $\widehat{\boldsymbol{\beta}}^w_{n(1*)}$ consists of non-zero elements of $\widehat{\boldsymbol{\beta}}^w_n$ and $\widehat{\boldsymbol{\beta}}^w_{n(2*)} = \mathbf{0}$. The asterisk here is to distinguish the partition of $\widehat{\boldsymbol{\beta}}^w_n$ from the partition of $\boldsymbol{\beta}_0$. If both partitions are the same, then the Weighted Lasso Bootstrap selects the true model. Based on the partition of $\widehat{\boldsymbol{\beta}}^w_n$, we have

$$\begin{aligned}
&2\sqrt{n}\left[C^w_n\left(\sqrt{n}\widehat{\boldsymbol{u}}_n\right) - \boldsymbol{Z}^w_n\right] \\
&= 2\sqrt{n}\left\{\begin{bmatrix} C^w_{n(11*)} & C^w_{n(12*)} \\ C^w_{n(21*)} & C^w_{n(22*)} \end{bmatrix} \times \sqrt{n}\begin{bmatrix} \widehat{\boldsymbol{u}}_{n(1*)} \\ \widehat{\boldsymbol{u}}_{n(2*)} \end{bmatrix} - \begin{bmatrix} \boldsymbol{Z}^w_{n(1*)} \\ \boldsymbol{Z}^w_{n(2*)} \end{bmatrix}\right\}.
\end{aligned}$$

As a consequence of the Karush-Kuhn-Tucker (KKT) conditions, we have

$$\begin{aligned}
&2\sqrt{n}\left\{\sqrt{n}\left[C^w_{n(11*)}\widehat{\boldsymbol{u}}_{n(1*)} + C^w_{n(12*)}\widehat{\boldsymbol{u}}_{n(2*)}\right] - \boldsymbol{Z}^w_{n(1*)}\right\} = -\lambda_n \mathrm{sgn}\left(\widehat{\boldsymbol{\beta}}^w_{n(1*)}\right) \\
&\implies C^w_{n(11*)}\left[\sqrt{n}\widehat{\boldsymbol{u}}_{n(1*)}\right] + C^w_{n(12*)}\left[\sqrt{n}\widehat{\boldsymbol{u}}_{n(2*)}\right] - \boldsymbol{Z}^w_{n(1*)} = -\frac{\lambda_n}{2\sqrt{n}}\mathrm{sgn}\left(\widehat{\boldsymbol{\beta}}^w_{n(1*)}\right),
\end{aligned} \tag{16}$$

and

$$\begin{aligned}
&\left|2\sqrt{n}\left\{\sqrt{n}\left[C^w_{n(21*)}\widehat{\boldsymbol{u}}_{n(1*)} + C^w_{n(22*)}\widehat{\boldsymbol{u}}_{n(2*)}\right] - \boldsymbol{Z}^w_{n(2*)}\right\}\right| \le \lambda_n \boldsymbol{J} \\
&\implies \left|C^w_{n(21*)}\left[\sqrt{n}\widehat{\boldsymbol{u}}_{n(1*)}\right] + C^w_{n(22*)}\left[\sqrt{n}\widehat{\boldsymbol{u}}_{n(2*)}\right] - \boldsymbol{Z}^w_{n(2*)}\right| \le \frac{\lambda_n}{2\sqrt{n}}\boldsymbol{J}
\end{aligned} \tag{17}$$

element wise. Hence, conditional on data, if there exists $\widehat{\boldsymbol{u}}_n$ such that the following equality and inequalities hold:

$$C^w_{n(11)}\left[\sqrt{n}\widehat{\boldsymbol{u}}_{n(1)}\right] - \boldsymbol{Z}^w_{n(1)} = -\frac{\lambda_n}{2\sqrt{n}}\mathrm{sgn}\left(\widehat{\boldsymbol{\beta}}^w_{n(1)}\right) \tag{18}$$

$$-\frac{\lambda_n}{2\sqrt{n}}\boldsymbol{J} \le C^w_{n(21)}\left[\sqrt{n}\widehat{\boldsymbol{u}}_{n(1)}\right] - \boldsymbol{Z}^w_{n(2)} \le \frac{\lambda_n}{2\sqrt{n}}\boldsymbol{J} \quad \text{element wise} \tag{19}$$

$$\left|\widehat{\boldsymbol{u}}_{n(1)}\right| < \left|\boldsymbol{\beta}_{0(1)}\right| \quad \text{element wise,} \tag{20}$$

then we have

$$\mathrm{sgn}\left[\widehat{\boldsymbol{\beta}}^w_{n(1)}\right] = \mathrm{sgn}\left[\boldsymbol{\beta}_{0(1)}\right] \qquad \text{and} \qquad \widehat{\boldsymbol{u}}_{n(2)} = \widehat{\boldsymbol{\beta}}^w_{n(2)} = \boldsymbol{\beta}_{0(2)} = \mathbf{0},$$

ie. $\widehat{\boldsymbol{\beta}}_n^w \overset{s}{=} \boldsymbol{\beta}_0$. This also implies that

$$P\left(\widehat{\boldsymbol{\beta}}_n^w \overset{s}{=} \boldsymbol{\beta}_0 \middle| \text{data}\right) \geq P\Bigg( \left\{ \left| C_{n(21)}^w \left[\sqrt{n}\widehat{\boldsymbol{u}}_{n(1)}\right] - \boldsymbol{Z}_{n(2)}^w \right| \leq \frac{\lambda_n}{2\sqrt{n}} \boldsymbol{J} \quad \text{element wise} \right\}$$
$$\bigcap \left\{ C_{n(11)}^w \left[\sqrt{n}\widehat{\boldsymbol{u}}_{n(1)}\right] - \boldsymbol{Z}_{n(1)}^w = -\frac{\lambda_n}{2\sqrt{n}} \text{sgn}\left(\widehat{\boldsymbol{\beta}}_{n(1)}^w\right) \right\}$$
$$\bigcap \left\{ \left|\widehat{\boldsymbol{u}}_{n(1)}\right| < \left|\boldsymbol{\beta}_{0(1)}\right| \quad \text{element wise} \right\} \Bigg| \text{data} \Bigg).$$

We now simplify the intersection of events on the RHS of the inequality. From (18),

$$\left(C_{n(11)}^w\right)^{-1} \boldsymbol{Z}_{n(1)}^w = \sqrt{n}\widehat{\boldsymbol{u}}_{n(1)} + \frac{\lambda_n}{2\sqrt{n}} \left(C_{n(11)}^w\right)^{-1} \text{sgn}\left(\boldsymbol{\beta}_{0(1)}\right)$$
$$\Longrightarrow \left| \left(C_{n(11)}^w\right)^{-1} \boldsymbol{Z}_{n(1)}^w \right| \leq \sqrt{n} \left|\widehat{\boldsymbol{u}}_{n(1)}\right| + \frac{\lambda_n}{2\sqrt{n}} \left| \left(C_{n(11)}^w\right)^{-1} \text{sgn}\left(\boldsymbol{\beta}_{0(1)}\right) \right|.$$

Substitute (20) into the inequality above yields

$$\left| \left(C_{n(11)}^w\right)^{-1} \boldsymbol{Z}_{n(1)}^w \right| \leq \sqrt{n} \left\{ \left|\boldsymbol{\beta}_{0(1)}\right| + \frac{\lambda_n}{2n} \left| \left(C_{n(11)}^w\right)^{-1} \text{sgn}\left(\boldsymbol{\beta}_{0(1)}\right) \right| \right\},$$

which corresponds to $A_n^w$. Also, from (18),

$$\sqrt{n}\widehat{\boldsymbol{u}}_{n(1)} = \left(C_{n(11)}^w\right)^{-1} \boldsymbol{Z}_{n(1)}^w - \frac{\lambda_n}{2\sqrt{n}} \left(C_{n(11)}^w\right)^{-1} \text{sgn}\left(\boldsymbol{\beta}_{0(1)}\right).$$

Substitute this equality into (19) yields

$$\left| C_{n(21)}^w \left(C_{n(11)}^w\right)^{-1} \boldsymbol{Z}_{n(1)}^w - \boldsymbol{Z}_{n(2)}^w - \frac{\lambda_n}{2\sqrt{n}} C_{n(21)}^w \left(C_{n(11)}^w\right)^{-1} \text{sgn}\left(\boldsymbol{\beta}_{0(1)}\right) \right| \leq \frac{\lambda_n}{2\sqrt{n}} \boldsymbol{J}$$

element wise, which can be expanded into

$$\left| C_{n(21)}^w \left(C_{n(11)}^w\right)^{-1} \boldsymbol{Z}_{n(1)}^w - \boldsymbol{Z}_{n(2)}^w - \frac{\lambda_n}{2\sqrt{n}} \left[ C_{n(21)} \left(C_{n(11)}\right)^{-1} \text{sgn}\left(\boldsymbol{\beta}_{0(1)}\right) + \boldsymbol{\rho}_n^w \right] \right| \leq \frac{\lambda_n}{2\sqrt{n}} \boldsymbol{J}$$

element wise. Now, from the Irrepresentable assumption,

$$\left| C_{n(21)} \left(C_{n(11)}\right)^{-1} \text{sgn}\left(\boldsymbol{\beta}_{0(1)}\right) \right| \leq \boldsymbol{J} - \boldsymbol{\eta}$$

$$\Longrightarrow \left|C_{n(21)}\left(C_{n(11)}\right)^{-1}\operatorname{sgn}\left(\boldsymbol{\beta}_{0(1)}\right)+\boldsymbol{\rho}_n^w\right| \leq \boldsymbol{J}-\left(\boldsymbol{\eta}-|\boldsymbol{\rho}_n^w|\right).$$

element wise. Consequently, we must have

$$\left|C_{n(21)}^w\left(C_{n(11)}^w\right)^{-1}\boldsymbol{Z}_{n(1)}^w-\boldsymbol{Z}_{n(2)}^w\right| \leq \frac{\lambda_n}{2\sqrt{n}}\left(\boldsymbol{\eta}-|\boldsymbol{\rho}_n^w|\right)$$

element wise, which corresponds to $B_n^w$. Therefore,

$$P\left(\widehat{\boldsymbol{\beta}}_n^w(\lambda_n) \overset{s}{=} \boldsymbol{\beta}_0\middle|\text{data}\right) \geq P\left(A_n^w \cap B_n^w\middle|\text{data}\right).$$

□

**Lemma 3.7.** *Assume all conditions stated in Theorem 2.3. Then*

$$P\left[(A_n^w)^c\,\middle|\,data\right]=o\left(e^{-n}\right).$$

*Proof.* From Lemma 3.1, we have

$$\frac{1}{n}X_{(1)}'D_nX_{(1)}=C_{n(11)}^w \xrightarrow{\text{a.s.}} C_{11}.$$

From Lemmas 3.3 and 3.4, we have

$$\frac{1}{\sqrt{n}}X_{(1)}'(D_n-I)\boldsymbol{\epsilon}_n\bigg|\text{data} \xrightarrow{\text{c.d.}} N\left(\mathbf{0},\sigma^2C_{11}\right)$$

$$\Longrightarrow \frac{1}{\sqrt{n}}X_{(1)}'D_n\boldsymbol{\epsilon}_n\bigg|\text{data}=\boldsymbol{Z}_{n(1)}^w|\text{data} \overset{d}{\approx} N\left(\frac{1}{\sqrt{n}}X_{(1)}'\boldsymbol{\epsilon}_n,\sigma^2C_{11}\right)$$

$$\Longrightarrow \left(C_{n(11)}^w\right)^{-1}\boldsymbol{Z}_{n(1)}^w\bigg|\text{data} \overset{d}{\approx} N\left[\frac{1}{\sqrt{n}}C_{11}^{-1}X_{(1)}'\boldsymbol{\epsilon}_n,\sigma^2C_{11}^{-1}\right].$$

For brevity, we introduce the following notations:

$$\begin{aligned}\boldsymbol{\kappa}_n &= [\kappa_{n,1}\dots\kappa_{n,q}]' \equiv \left(C_{n(11)}^w\right)^{-1}\boldsymbol{Z}_{n(1)}^w\\ \boldsymbol{\mu}_n &= [\mu_{n,1}\dots\mu_{n,q}]' \equiv \frac{1}{\sqrt{n}}\left(C_{11}\right)^{-1}X_{(1)}'\boldsymbol{\epsilon}_n\\ \boldsymbol{b}_n &= [b_{n,1}\dots b_{n,q}]' \equiv \left(C_{n(11)}^w\right)^{-1}\operatorname{sgn}\left(\boldsymbol{\beta}_{0(1)}\right).\end{aligned}$$

It is easy to note that $\boldsymbol{b}_n=\mathcal{O}(1)$ since

$$\boldsymbol{b}_n \to C_{11}^{-1}\operatorname{sgn}\left(\boldsymbol{\beta}_{0(1)}\right).$$

In addition, by Lemma 3.1 of Chatterjee and Lahiri (2011),

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\boldsymbol{\mu}_n = (C_{11})^{-1}\frac{1}{n}X'_{(1)}\boldsymbol{\epsilon}_n \xrightarrow{\text{a.s.}} \mathbf{0} \\
\Longrightarrow\ &\frac{1}{\sqrt{n}}\boldsymbol{\mu}_n\bigg|\text{data} \xrightarrow{\text{c.p.}} \mathbf{0}.
\end{aligned}$$

Furthermore, from above, it is clear that each of $\kappa_{n,j}$ for $j = 1, \ldots, q$ has a normal limiting distribution with finite variance $\sigma^2_{\kappa,j}$. Then there exists $\sigma^2_\kappa$ such that

$$\sigma^2_{\kappa,j} \leq \sigma^2_\kappa \quad \forall \quad j = 1, \ldots, q,$$

and let

$$\tau_{n,j} = \frac{\kappa_{n,j} - \mu_{n,j}}{\sigma_{\kappa,j}},$$

which has a standard Normal limiting distribution. Since

$$(A^w_n)^c = \left\{\left|\left(C^w_{n(11)}\right)^{-1}\boldsymbol{Z}^w_{n(1)}\right| \geq \sqrt{n}\left[\left|\boldsymbol{\beta}_{0(1)}\right| - \frac{\lambda_n}{2n}\left|\left(C^w_{n(11)}\right)^{-1}\text{sgn}\left(\boldsymbol{\beta}_{0(1)}\right)\right|\right] \text{ element-wise}\right\},$$

then it follows that

$$\begin{aligned}
P\left[(A^w_n)^c\bigg|\text{data}\right] &= P\left[\bigcap_{j=1}^{q}\left\{|\kappa_{n,j}| \geq \sqrt{n}\left(|\beta_{0,j}| + \frac{\lambda_n}{2n}b_{n,j}\right)\right\}\bigg|\text{data}\right] \\
&\leq \sum_{j=1}^{q} P\left[|\kappa_{n,j}| \geq \sqrt{n}\left(|\beta_{0,j}| + \frac{\lambda_n}{2n}b_{n,j}\right)\bigg|\text{data}\right] \\
&= \sum_{j=1}^{q} P\left(\tau_{n,j} \geq \frac{\sqrt{n}}{\sigma_{\kappa,j}}\left[|\beta_{0,j}| + \frac{\lambda_n}{2n}b_{n,j} - \frac{1}{\sqrt{n}}\mu_{n,j}\right]\bigg|\text{data}\right) \\
&\quad + \sum_{j=1}^{q} P\left(\tau_{n,j} \leq -\frac{\sqrt{n}}{\sigma_{\kappa,j}}\left[|\beta_{0,j}| + \frac{\lambda_n}{2n}b_{n,j} + \frac{1}{\sqrt{n}}\mu_{n,j}\right]\bigg|\text{data}\right) \\
&\leq 2\sum_{j=1}^{q}\left[1 - \Phi\left\{\frac{|\beta_{0,j}| + o(1)}{\sigma_\kappa/\sqrt{n}}\right\}\right] \\
&\leq 2\sum_{j=1}^{q}\frac{\sigma_\kappa}{[1 + o(1)]\sqrt{n}|\beta_{0,j}|}\exp\left\{-\frac{n\left[\beta_{0,j} + o(1)\right]^2}{2\sigma^2_\kappa}\right\} \\
&= o\left(e^{-n}\right),
\end{aligned}$$

where the second last line follows from the well-known Gaussian tail bound

$$1 - \Phi(t) \leq t^{-1} e^{-t^2}. \tag{21}$$

□

**Lemma 3.8.** *Assume all conditions stated in Theorem 2.3. Denote*

$$(C_n^{w*})' \equiv \left[ \left( \frac{1}{n} X'_{(2)} D_n X_{(2)} \right) \left( \frac{1}{n} X'_{(1)} D_n X_{(1)} \right)^{-1} X'_{(1)} - X'_{(2)} \right].$$

*Then*

$$\frac{1}{\sqrt{n}} (C_n^{w*})' (D_n - I) \boldsymbol{\epsilon}_n \bigg| data \xrightarrow{c.d.} N(\mathbf{0}, \Sigma),$$

*where* $\Sigma = \sigma^2 \left( C_{22} - C_{21} C_{11}^{-1} C'_{21} \right)$.

*Proof.* Let $\left( \boldsymbol{c}_{n,i}^{w*} \right)$ be the $i$-th row of the matrix $C_n^{w*}$, so that

$$\begin{aligned}
&\frac{1}{\sqrt{n}} (C_n^{w*})' (D_n - I) \boldsymbol{\epsilon}_n \\
&= \left[ \frac{1}{n} \sum_{i=1}^n \epsilon_i^2 \left( \boldsymbol{c}_{n,i}^{w*} \right) \left( \boldsymbol{c}_{n,i}^{w*} \right)' \right]^{\frac{1}{2}} \times \left[ \frac{1}{n} \sum_{i=1}^n \epsilon_i^2 \left( \boldsymbol{c}_{n,i}^{w*} \right) \left( \boldsymbol{c}_{n,i}^{w*} \right)' \right]^{-\frac{1}{2}} \sum_{i=1}^n \left( \boldsymbol{c}_{n,i}^{w*} \right) \epsilon_i (W_i - 1).
\end{aligned}$$

First, we want to show that

$$\frac{1}{n} \sum_{i=1}^n \epsilon_i^2 \left( \boldsymbol{c}_{n,i}^{w*} \right) \left( \boldsymbol{c}_{n,i}^{w*} \right)' \bigg| \text{data} \xrightarrow{\text{c.p.}} \Sigma.$$

We begin with

$$\begin{aligned}
&\frac{1}{n} \sum_{i=1}^n \left( \boldsymbol{c}_{n,i}^{w*} \right) \left( \boldsymbol{c}_{n,i}^{w*} \right)' \\
&= \frac{1}{n} (C_n^{w*})' (C_n^{w*}) \\
&= \frac{1}{n} \left[ C_{n(21)}^w \left( C_{n(11)}^w \right)^{-1} X'_{(1)} - X'_{(2)} \right] \left[ X_{(1)} \left( C_{n(11)}^w \right)^{-1} \left( C_{n(21)}^w \right)' - X_{(2)} \right] \\
&\xrightarrow{\text{a.s.}} C_{21} C_{11}^{-1} C_{11} C_{11}^{-1} C'_{21} + C_{22} - 2 C_{21} C_{11}^{-1} C'_{21} \\
&= C_{22} - C_{21} C_{11}^{-1} C'_{21}.
\end{aligned}$$

By Lemma 3.1 and assumption (3),

$$\underset{1\leq i\leq n}{max}\left\|\boldsymbol{c}_{n,i}^{w*}\right\|_2^2=\mathcal{O}(1)\quad\text{a.s.}$$

Then, by Strong Law of Large Numbers,

$$\frac{1}{n}\sum_{i=1}^{n}\left(\epsilon_i^2-\sigma^2\right)\left(\boldsymbol{c}_{n,i}^{w*}\right)\left(\boldsymbol{c}_{n,i}^{w*}\right)'\xrightarrow{\text{a.s.}}\mathbf{0}.$$

Hence, by Continuous Mapping Theorem,

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2\left(\boldsymbol{c}_{n,i}^{w*}\right)\left(\boldsymbol{c}_{n,i}^{w*}\right)'\xrightarrow{\text{a.s.}}\sigma^2\left[C_{22}-C_{21}C_{11}^{-1}C_{21}'\right]=\Sigma\\ \Longrightarrow\;&\frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2\left(\boldsymbol{c}_{n,i}^{w*}\right)\left(\boldsymbol{c}_{n,i}^{w*}\right)'\Big|\text{data}\xrightarrow{\text{c.p.}}\Sigma.\end{aligned}$$

For the second part of the proof, we want to show that Lindeberg's Central Limit Theorem gives

$$\left[\frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2\left(\boldsymbol{c}_{n,i}^{w*}\right)\left(\boldsymbol{c}_{n,i}^{w*}\right)'\right]^{-\frac{1}{2}}\left[\sum_{i=1}^{n}\left(\boldsymbol{c}_{n,i}^{w*}\right)\epsilon_i(W_i-1)\right]\Bigg|\text{data}\xrightarrow{\text{c.d.}}N\left(\mathbf{0},I\right).$$

Without loss of generality, we show that the Liapounov's sufficient condition is satisfied for the univariate case:

$$\sum_{i=1}^{n}9\epsilon_i^4\left(c_{n,i}^{w*}\right)^4=o\left(\left[\sum_{i=1}^{n}\epsilon_i^2\left(c_{n,i}^{w*}\right)^2\right]^2\right)\quad\text{given data.}$$

By Lemma 3.1 and assumptions (3) and (8),

$$\sum_{i=1}^{n}9\epsilon_i^4\left(c_{n,i}^{w*}\right)^4=\mathcal{O}(n)\quad\text{given data.}$$

Based on the first part of the proof, we can see that

$$\left[\sum_{i=1}^{n}\epsilon_i^2\left(c_{n,i}^{w*}\right)^2\right]^2=\mathcal{O}(n^2)\quad\text{given data.}$$

Thus Liapounov's sufficient condition is satisfied. Finally, by Lemma 2.1,

$$\frac{1}{\sqrt{n}}\left(C_n^{w*}\right)'(D_n - I)\boldsymbol{\epsilon}_n\Big|\text{data} \xrightarrow{\text{c.d.}} N\left(\mathbf{0}, \Sigma\right).$$

□

**Lemma 3.9.** *Assume all conditions stated in Theorem 2.3. Then*

$$P\left[(B_n^w)^c \,\middle|\, data\right] = o\left(e^{-n^{2c-1}}\right).$$

*Proof.* First note that

$$(B_n^w)^c = \left\{\left|C_{n(21)}^w \left(C_{n(11)}^w\right)^{-1} \boldsymbol{Z}_{n(1)}^w - \boldsymbol{Z}_{n(2)}^w\right| > \frac{\lambda_n}{2\sqrt{n}}\left(\boldsymbol{\eta} - |\boldsymbol{\rho}_n^w|\right) \text{ element-wise}\right\}.$$

By Lemma 3.1 and assumption 4,

$$\begin{aligned}\boldsymbol{\rho}_n^w &= \left[C_{n(21)}^w \left(C_{n(11)}^w\right)^{-1} - C_{n(21)}\left(C_{n(11)}\right)^{-1}\right] \operatorname{sgn}\left(\boldsymbol{\beta}_{0(1)}\right) \\ &\xrightarrow{\text{a.s.}} \left[C_{21}\left(C_{11}\right)^{-1} - C_{21}\left(C_{11}\right)^{-1}\right] \operatorname{sgn}\left(\boldsymbol{\beta}_{0(1)}\right) \\ &= \mathbf{0},\end{aligned}$$

and hence, $\boldsymbol{\rho}_n^w = o(1)$ a.s. Next, using the notations from Lemma 3.8,

$$\begin{aligned}&C_{n(21)}^w \left(C_{n(11)}^w\right)^{-1} \boldsymbol{Z}_{n(1)}^w - \boldsymbol{Z}_{n(2)}^w \\ &= \frac{1}{\sqrt{n}}\left[\left(\frac{1}{n}X_{(2)}' D_n X_{(2)}\right)\left(\frac{1}{n}X_{(1)}' D_n X_{(1)}\right)^{-1} X_{(1)}' - X_{(2)}'\right] D_n \boldsymbol{\epsilon}_n \\ &= \frac{1}{\sqrt{n}}\left(C_n^{w*}\right)' D_n \boldsymbol{\epsilon}_n.\end{aligned}$$

From Lemma 3.8,

$$\begin{aligned}&\frac{1}{\sqrt{n}}\left(C_n^{w*}\right)'(D_n - I)\boldsymbol{\epsilon}_n\Big|\text{data} \xrightarrow{\text{c.d.}} N\left(\mathbf{0}, \Sigma\right) \\ &\implies \frac{1}{\sqrt{n}}\left(C_n^{w*}\right)' D_n \boldsymbol{\epsilon}_n\Big|\text{data} \overset{d}{\approx} N\left(C_{22}C_{11}^{-1}\frac{1}{\sqrt{n}}X_{(1)}'\boldsymbol{\epsilon}_n - \frac{1}{\sqrt{n}}X_{(2)}'\boldsymbol{\epsilon}_n, \Sigma\right).\end{aligned}$$

Again, for brevity, we introduce the following notations:

$$\boldsymbol{\xi}_n = \left[\xi_{n,1}\ldots\xi_{n,p-q}\right]' \equiv C_{n(21)}^w \left(C_{n(11)}^w\right)^{-1} \boldsymbol{Z}_{n(1)}^w - \boldsymbol{Z}_{n(2)}^w$$

$$\boldsymbol{\nu}_n = [\nu_{n,1} \dots \nu_{n,p-q}]' \equiv C_{22}C_{11}^{-1}\frac{1}{\sqrt{n}}X'_{(1)}\boldsymbol{\epsilon}_n - \frac{1}{\sqrt{n}}X'_{(2)}\boldsymbol{\epsilon}_n.$$

By Lemma 3.2 of Chatterjee and Lahiri (2011),

$$\begin{aligned} &\frac{1}{n^{c-\frac{1}{2}}}\boldsymbol{\nu}_n \\ &= C_{21}C_{11}^{-1}\frac{1}{n^c}X'_{(1)}\boldsymbol{\epsilon}_n - \frac{1}{n^c}X'_{(2)}\boldsymbol{\epsilon}_n \\ &\xrightarrow{\text{a.s.}} \mathbf{0}. \end{aligned}$$

Besides that, note that each of $\xi_{n,j}$ for $j = 1, \dots, p-q$ has a Normal limiting distribution with finite variance $\sigma^2_{\xi,j}$. Then, there exists $\sigma^2_\xi$ such that

$$\sigma^2_{\xi,j} \le \sigma^2_\xi \quad \forall \quad j = 1, \dots, p-q,$$

and let

$$\zeta_{n,j} = \frac{\xi_{n,j} - \nu_{n,j}}{\sigma_{\xi,j}},$$

which has a standard Normal limiting distribution. Therefore,

$$\begin{aligned} P\left[(B_n^w)^c \,\middle|\text{data}\right] &= P\left(\bigcap_{j=1}^{p-q}\left\{|\xi_{n,j}| > \frac{\lambda_n}{2\sqrt{n}}\left(\eta_j - \left|\rho^w_{n,j}\right|\right)\right\}\middle|\text{data}\right) \\ &\le \sum_{j=1}^{p-q} P\left(|\xi_{n,j}| > \frac{\lambda_n}{2\sqrt{n}}\left(\eta_j - \left|\rho^w_{n,j}\right|\right)\middle|\text{data}\right) \\ &= \sum_{j=1}^{p-q} P\left(\zeta_{n,j} > \frac{\frac{\lambda_n}{2\sqrt{n}}\left(\eta_j - \left|\rho^w_{n,j}\right|\right) - \nu_{n,j}}{\sigma_{\xi,j}}\,\middle|\,\text{data}\right) \\ &\quad + \sum_{j=1}^{p-q} P\left(\zeta_{n,j} < -\frac{\frac{\lambda_n}{2\sqrt{n}}\left(\eta_j - \left|\rho^w_{n,j}\right|\right) + \nu_{n,j}}{\sigma_{\xi,j}}\,\middle|\,\text{data}\right) \\ &= \sum_{j=1}^{p-q} P\left(\zeta_{n,j} > \frac{n^{c-\frac{1}{2}}}{2\sigma_{\xi,j}}\left[\frac{\lambda_n}{n^c}\eta_j - \frac{\lambda_n}{n^c}\left|\rho^w_{n,j}\right| - \frac{2\nu_{n,j}}{n^{c-\frac{1}{2}}}\right]\,\middle|\,\text{data}\right) \\ &\quad + \sum_{j=1}^{p-q} P\left(\zeta_{n,j} < -\frac{n^{c-\frac{1}{2}}}{2\sigma_{\xi,j}}\left[\frac{\lambda_n}{n^c}\eta_j - \frac{\lambda_n}{n^c}\left|\rho^w_{n,j}\right| + \frac{2\nu_{n,j}}{n^{c-\frac{1}{2}}}\right]\,\middle|\,\text{data}\right) \\ &\le 2\sum_{j=1}^{p-q}\left[1 - \Phi\left\{\frac{n^{c-\frac{1}{2}}}{2\sigma_\xi}\left[\frac{\lambda_n}{n^c}\eta_j + o(1)\right]\right\}\right] \end{aligned}$$

$$
\begin{aligned}
&\leq 2\sum_{j=1}^{p-q} \frac{2\sigma_\xi}{n^{c-\frac{1}{2}}\left[\frac{\lambda_n}{n^c}\eta_j + o(1)\right]} \exp\left\{-\frac{1}{2}\left(\frac{n^{c-\frac{1}{2}}}{2\sigma_\xi}\left[\frac{\lambda_n}{n^c}\eta_j + o(1)\right]\right)^2\right\} \\
&= o\left(e^{-n^{2c-1}}\right),
\end{aligned}
$$

where the second last line follows from (21). □

We are now ready to prove Theorem 2.3.

*Proof of Theorem 2.3.* From Lemma 3.6,

$$
\begin{aligned}
P\left(\widehat{\boldsymbol{\beta}}_n^w(\lambda_n) \overset{s}{=} \boldsymbol{\beta}_0 \middle| \text{data}\right) &\geq P\left(A_n^w \bigcap B_n^w \middle| \text{data}\right) \\
&= 1 - P\left[\left(A_n^w \bigcap B_n^w\right)^c \middle| \text{data}\right] \\
&= 1 - P\left[(A_n^w)^c \bigcup (B_n^w)^c \middle| \text{data}\right] \\
&\geq 1 - \left\{P\left[(A_n^w)^c \middle| \text{data}\right] + P\left[(B_n^w)^c \middle| \text{data}\right]\right\} \\
&= 1 - o\left(e^{-n^{2c-1}}\right),
\end{aligned}
$$

where the last line follows from Lemmas 3.7 and 3.9. □

# References

Chatterjee, A. and Lahiri, S. N. (2011), "Strong consistency of lasso estimators," *Sankhya: The Indian Journal of Statistics, Series A*, 73, 55–78.

Geyer, C. (1996), "On the asymptotics of convex stochastic optimization," Unpublished manuscript.

Knight, K. and Fu, W. (2000), "Asymptotics for lasso-type estimators," *The Annals of Statistics*, 28, 1356–1378.

Lai, T. L., Robbins, H., and Wei, C. Z. (1978), "Strong consistency of least squares estimates in multiple regression," *Proceedings of National Academy of Sciences*, 75, 3034 – 3036.

Newton, M. A. (1991), "The weighted likelihood bootstrap and an algorithm for prepivoting," Ph.D. thesis, University of Washington, Seattle.

Newton, M. A. and Raftery, A. (1994), "Approximate bayesian inference with the weighted likelihood bootstrap," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 56, 3–48.

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

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

Zhao, P. and Yu, B. (2006), "On model selection consistency of Lasso," *Journal of Machine Learning Research*, 7, 2541–2563.