Technical Proofs for "Homogeneity Pursuit"

Abstract

This is the supplemental material for the article "Homogeneity Pursuit", submitted for publication in Journal of the American Statistical Association.

B Proofs

B.1 Proof of Theorem 3.5

Since τ is consistent with groups in $\boldsymbol{\beta}^0$, there exists $1 = j_1 < j_2 < \cdots < j_{K+1} = p+1$ such that $A_k = \{\tau(j_k), \tau(j_k+1), \cdots, \tau(j_{k+1}-1)\}$ for all k. We shall write $\tau(j) = j$ without loss of generality.

In the first part of the proof, we show that $\widehat{\boldsymbol{\beta}} \in \mathcal{M}_A$, and it satisfies the sign restrictions $\mathrm{sgn}(\widehat{\beta}_{A,k+1} - \widehat{\beta}_{A,k}) = \mathrm{sgn}(\beta^0_{A,k+1} - \beta^0_{A,k})$, $k = 1, \cdots, K-1$.

When $\rho(t) = |t|$, $Q_n(\boldsymbol{\beta})$ is strictly convex. So $\widehat{\boldsymbol{\beta}}$ is the unique global minimizer if and only if it satisfies the first-order conditions:

$$0 = \begin{cases} -\frac{1}{n}\mathbf{x}_1^T\boldsymbol{\varepsilon} + \frac{1}{n}\mathbf{x}_1^T\mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) - \lambda_n \mathrm{sgn}(\widehat{\beta}_2 - \widehat{\beta}_1), \\ -\frac{1}{n}\mathbf{x}_j^T\boldsymbol{\varepsilon} + \frac{1}{n}\mathbf{x}_j^T\mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) + \lambda_n \mathrm{sgn}(\widehat{\beta}_j - \widehat{\beta}_{j-1}) - \lambda_n \mathrm{sgn}(\widehat{\beta}_{j+1} - \widehat{\beta}_j), \quad 2 \le j \le p \\ -\frac{1}{n}\mathbf{x}_p^T\boldsymbol{\varepsilon} + \frac{1}{n}\mathbf{x}_p^T\mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) + \lambda_n \mathrm{sgn}(\widehat{\beta}_p - \widehat{\beta}_{p-1}), \end{cases}$$

where $\mathrm{sgn}(t) = 1$ when $t > 0$, -1 when $t < 0$, and any value in $[-1, 1]$ when $t = 0$. Therefore, it suffices to show that there exists $\widehat{\boldsymbol{\beta}} \in \mathcal{M}_A$ that satisfies the sign restrictions and the first-order conditions simultaneously.

For $\widehat{\boldsymbol{\beta}} \in \mathcal{M}_A$, we write $\widehat{\boldsymbol{\mu}} = T(\widehat{\boldsymbol{\beta}})$ and $\boldsymbol{\mu}^0 = T(\boldsymbol{\beta}^0)$, where the mapping T is the same as that in the proof of Theorem 3.1. The sign restrictions now become $\mathrm{sgn}(\widehat{\mu}_{k+1} - \widehat{\mu}_k) = \mathrm{sgn}(\mu^0_{k+1} - \mu^0_k)$ for all $k = 1, \cdots, K-1$. Note that $\widehat{\beta}_j = \widehat{\beta}_{j+1}$ when predictors j and

$(j+1)$ belong to the same group in $\mathcal{A}$. The first-order conditions can be re-expressed as

$$0 = \begin{cases} -\frac{1}{n}\mathbf{x}_j^T\boldsymbol{\varepsilon} + \frac{1}{n}\mathbf{x}_j^T\mathbf{X}_A(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}^0) + \lambda_n \mathrm{sgn}(\widehat{\mu}_k - \widehat{\mu}_{k-1}) - \lambda_n r_j, & j = j_k \\ -\frac{1}{n}\mathbf{x}_j^T\boldsymbol{\varepsilon} + \frac{1}{n}\mathbf{x}_j^T\mathbf{X}_A(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}^0) + \lambda_n r_{j-1} - \lambda_n \mathrm{sgn}(\widehat{\mu}_{k+1} - \widehat{\mu}_k), & j = j_{k+1} - 1 \\ -\frac{1}{n}\mathbf{x}_j^T\boldsymbol{\varepsilon} + \frac{1}{n}\mathbf{x}_j^T\mathbf{X}_A(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}^0) + \lambda_n r_{j-1} - \lambda_n r_j, & \text{elsewhere,} \end{cases} \tag{1}$$

where r_j's take any values on $[-1, 1]$ and we set $\mathrm{sgn}(\widehat{\mu}_1 - \widehat{\mu}_0) = \mathrm{sgn}(\widehat{\mu}_{K+1} - \widehat{\mu}_K) = 0$ by default. Denote by $\delta_k^0 = \mathrm{sgn}(\mu_{k+1}^0 - \mu_k^0)$ when $1 \le k \le K-1$ and $\delta_k^0 = 0$ when $k = 0, K$; similarly, $\widehat{\delta}_k$ for $1 \le k \le K$. In (1), we first remove r_j's by summing up the equations corresponding to indices in each A_k. Using the fact that $\mathbf{x}_{A,k} = \sum_{j \in A_k} \mathbf{x}_j$, we obtain

$$-\tfrac{1}{n}\mathbf{x}_{A,k}^T\boldsymbol{\varepsilon} + \tfrac{1}{n}\mathbf{x}_{A,k}^T\mathbf{X}_A(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}^0) + \lambda_n\widehat{\delta}_{k-1} - \lambda_n\widehat{\delta}_k = 0, \quad k = 1, \cdots, K.$$

Under the sign restrictions $\widehat{\delta}_k = \delta_k^0$, $k = 1, \cdots, K-1$, it becomes a pure linear equation of $(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}^0)$:

$$-\tfrac{1}{n}\mathbf{X}_A^T\boldsymbol{\varepsilon} + \tfrac{1}{n}\mathbf{X}_A^T\mathbf{X}_A(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}^0) - \lambda_n\mathbf{d}^0 = 0,$$

where $\mathbf{d}^0$ is the K-dimensional vector with $d_k^0 = \delta_k^0 - \delta_{k-1}^0$, as defined in Section 3.4. It follows immediately that

$$\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}^0 = n\lambda_n(\mathbf{X}_A^T\mathbf{X}_A)^{-1}\mathbf{d}^0 + (\mathbf{X}_A^T\mathbf{X}_A)^{-1}\mathbf{X}_A^T\boldsymbol{\varepsilon}. \tag{2}$$

Second, given $(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}^0)$, (1) can be viewed as equations of r_j's and we can solve them directly. Denote $\boldsymbol{\xi} = \frac{1}{n}\mathbf{X}^T\mathbf{X}_A(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}^0) - \frac{1}{n}\mathbf{X}^T\boldsymbol{\varepsilon}$. For each $j \in A_k$, define $A_{kj}^1 = \{j_k, \cdots, j\}$ and $A_{kj}^2 = \{j+1, \cdots, j_{k+1} - 1\}$. The solutions of (1) are

$$r_j = \widehat{\delta}_{k-1} + \lambda_n^{-1}\sum_{i \in A_{kj}^1}\xi_i = \widehat{\delta}_k - \lambda_n^{-1}\sum_{i \in A_{kj}^2}\xi_i, \qquad j \in A_k, \ \ j \neq j_{k+1} - 1.$$

Here the two expressions of r_j are equivalent because $\lambda_n\sum_{i \in A_k}\xi_i = \widehat{\delta}_k - \widehat{\delta}_{k-1}$ from (1). It follows that any convex combination of the two expressions is also an equivalent expression of r_j. Taking the combination coefficients as $|A_{kj}^2|/|A_k|$ and $|A_{kj}^1|/|A_k|$, and plugging in the sign restrictions $\widehat{\delta}_k = \delta_k^0$, $k = 1, \cdots, K-1$, we obtain

$$\begin{aligned} r_j &= \lambda_n^{-1}\Big(\frac{|A_{kj}^2|}{|A_k|}\sum_{i \in A_{kj}^1}\xi_i - \frac{|A_{kj}^1|}{|A_k|}\sum_{i \in A_{kj}^2}\xi_i\Big) + \Big(\frac{|A_{kj}^2|}{|A_k|}\delta_{k-1}^0 + \frac{|A_{kj}^1|}{|A_k|}\delta_k^0\Big) \\ &= n\lambda_n^{-1}w_j(\boldsymbol{\xi}) + \Big(\frac{|A_{kj}^2|}{|A_k|}\delta_{k-1}^0 + \frac{|A_{kj}^1|}{|A_k|}\delta_k^0\Big), \end{aligned}$$

where the function $w_j(\cdot)$ is defined as in (36). Here r_j's still depend on $(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}^0)$ through $\boldsymbol{\xi}$. Combining (2) to the definition of $\boldsymbol{\xi}$ gives

$$\begin{aligned}\boldsymbol{\xi} &= -\tfrac{1}{n}\mathbf{X}^T\left[\mathbf{I} - \mathbf{X}_A(\mathbf{X}_A^T\mathbf{X}_A)^{-1}\mathbf{X}_A^T\right]\boldsymbol{\varepsilon} + \lambda_n \mathbf{X}^T\mathbf{X}_A(\mathbf{X}_A^T\mathbf{X}_A)^{-1}\mathbf{d}^0\\ &\equiv -\tfrac{1}{n}\mathbf{X}^T\bar{\mathbf{P}}_A\boldsymbol{\varepsilon} + \lambda_n\mathbf{b}^0,\end{aligned}$$

where $\bar{\mathbf{P}}_A = \mathbf{I} - \mathbf{X}_A(\mathbf{X}_A^T\mathbf{X}_A)^{-1}\mathbf{X}_A^T$ and $\mathbf{b}^0$ is defined as in Section 3.4. By plugging in the expression of $\boldsymbol{\xi}$, we can remove the dependence on $(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}^0)$ of the solutions r_j's:

$$r_j = -\lambda_n^{-1} w_j(\mathbf{X}^T\bar{\mathbf{P}}_A\boldsymbol{\varepsilon}) + n w_j(\mathbf{b}^0) + \Big(\frac{|A_{kj}^2|}{|A_k|}\delta_{k-1}^0 + \frac{|A_{kj}^1|}{|A_k|}\delta_k^0\Big). \tag{3}$$

Now, to show the existence of $\widehat{\boldsymbol{\beta}} \in \mathcal{M}_A$ that satisfies both the sign restrictions and first-order conditions, it suffices to show with probability at least $1 - \epsilon_0 - n^{-1}K - (n \vee p)^{-1}$,

(a) the r_j's in (3) take values on $[-1, 1]$;

(b) the $\widehat{\boldsymbol{\mu}}$ in (2) satisfy the sign restrictions, i.e., $\mathrm{sgn}(\widehat{\mu}_{k+1} - \widehat{\mu}_k) = \mathrm{sgn}(\mu_{k+1}^0 - \mu_k^0)$ for all $k = 1, \cdots, K-1$.

Consider (a) first. In (3), by Condition 3.4, the sum of the last two terms is bounded by $(1 - \omega_n)$ in magnitude. To deal with the first term, recall that in deriving (38), we write $w_j(\mathbf{X}^T\boldsymbol{\varepsilon}) = \mathbf{a}_j^T\boldsymbol{\varepsilon}$. It follows immediately that $w_j(\mathbf{X}^T\bar{\mathbf{P}}_A\boldsymbol{\varepsilon}) = \mathbf{a}_j^T\bar{\mathbf{P}}_A\boldsymbol{\varepsilon} = (\bar{\mathbf{P}}_A\mathbf{a}_j)^T\boldsymbol{\varepsilon}$. Since $\|\bar{\mathbf{P}}_A\mathbf{a}_j\| \le \|\mathbf{a}_j\|$, similarly to (38), we obtain

$$\max_{j \in A_k} |w_j(\mathbf{X}^T\bar{\mathbf{P}}_A\boldsymbol{\varepsilon})| \le C\sqrt{\sigma_k|A_k|\log(n \vee p)/n}, \qquad 1 \le k \le K,$$

except for a probability at most $(n \vee p)^{-1}$. Therefore, by the choice of λ_n in (main-18), the absolute value of the first term is much smaller than ω_n. So $\max_j |r_j| \le 1$ except for a probability at most $(n \vee p)^{-1}$, i.e., (a) holds.

Next, consider (b). Since $|\mu_{k+1}^0 - \mu_k^0| \ge 2b_n$, it suffices to show that $\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}^0\|_\infty < b_n$. Note that (2) can be rewritten as

$$\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}^0 = \mathbf{D}^{-1}(\tfrac{1}{n}\mathbf{D}^{-1}\mathbf{X}_A^T\mathbf{X}_A\mathbf{D}^{-1})^{-1}(\lambda_n\mathbf{D}^{-1}\mathbf{d}^0 + n^{-1}\mathbf{D}^{-1}\mathbf{X}_A^T\boldsymbol{\varepsilon}).$$

It follows from Condition 3.1 that $\|\boldsymbol{\mu} - \boldsymbol{\mu}^0\| \le c_1^{-1}(\lambda_n\|\mathbf{D}^{-2}\mathbf{d}^0\| + n^{-1}\|\mathbf{D}^{-1}\|\|\mathbf{D}^{-1}\mathbf{X}_A^T\boldsymbol{\varepsilon}\|)$. First, $\|\mathbf{D}^{-2}\mathbf{d}^0\|^2 \le 4\sum_{k=1}^K \frac{1}{|A_k|^2}$. Second, from (26), $\|\mathbf{D}^{-1}\mathbf{X}_A^T\boldsymbol{\varepsilon}\| \le C\sqrt{nK\log(n)}$, except a probability of at most $n^{-1}K$. Moreover, $\|\mathbf{D}^{-1}\| = (\min_k |A_k|)^{-1/2} \le 1$. These together imply

$$\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}^0\| \le C\lambda_n\Big(\sum_{k=1}^K \frac{1}{|A_k|^2}\Big)^{1/2} + C\sqrt{\frac{K\log(n)}{n}}.$$

From (main-18), the right hand side is much smaller than b_n. It follows that $\|\widehat{\boldsymbol{\mu}}-\boldsymbol{\mu}^0\|_\infty \ll b_n$. This proves (b).

In the second part of the proof, we derive the convergence rate of $\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\|$. Note that $\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\| = \|\mathbf{D}(\widehat{\boldsymbol{\mu}}-\boldsymbol{\mu}^0)\|$, and from (2),

$$\mathbf{D}(\widehat{\boldsymbol{\mu}}-\boldsymbol{\mu}^0) = (\tfrac{1}{n}\mathbf{D}^{-1}\mathbf{X}_A^T\mathbf{X}_A\mathbf{D}^{-1})^{-1}\big(\lambda_n\mathbf{D}^{-1}\mathbf{d}^0 + n^{-1}\mathbf{D}^{-1}\mathbf{X}_A^T\boldsymbol{\varepsilon}\big).$$

Therefore, $\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\| \leq c_1^{-1}(\lambda_n\|\mathbf{D}^{-1}\mathbf{d}^0\|+n^{-1}\|\mathbf{D}^{-1}\mathbf{X}_A^T\boldsymbol{\varepsilon}\|)$, where $\|\mathbf{D}^{-1}\mathbf{d}^0\|^2 \leq 4\sum_{k=1}^K \frac{1}{|A_k|}$ and $\|\mathbf{D}^{-1}\mathbf{X}_A^T\boldsymbol{\varepsilon}\| = O_p(\sqrt{nK})$ by (24). Combining these gives

$$\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\| = O_p\Big(\sqrt{K/n} + \lambda_n\big(\sum_k \tfrac{1}{|A_k|}\big)^{1/2}\Big).$$

□

B.2 A useful proposition and its proof

The requirement that Υ preserves the order of $\boldsymbol{\beta}^0$ implies restrictions on how much the ordering (in terms of increasing values) of coordinates in $\widetilde{\boldsymbol{\beta}}$ deviates from that of $\boldsymbol{\beta}^0$. This is reflected on how the segments $\{B_1,\cdots,B_L\}$ intersect with the true groups $\{A_1,\cdots,A_K\}$. Recall that $V_{kl} = A_k \cap B_l$. We have the following proposition:

Proposition B.1. *When Υ preserves the order of $\boldsymbol{\beta}^0$, for each k, there exist d_k and u_k such that $A_k = \cup_{d_k\leq l\leq u_k} V_{kl}$, and $V_{kl} = B_l$ for $d_k < l < u_k$. For each l, there exist a_l and b_l such that $B_l = \cup_{a_l\leq k\leq b_l} V_{kl}$, and $V_{kl} = A_k$ for $a_l < k < b_l$.*

Proposition B.1 indicates that there are two cases for each A_k: either A_k is contained in a single B_l or it is contained in some consecutive B_l's where except the first and last ones, all the other B_l's are fully occupied by A_k. Similarly, there are two cases for each B_l: either it is contained in a single A_k or it is contained in some consecutive A_k's where except the first and last ones, all the other A_k's are fully occupied by B_l.

Proof. Consider the first claim. Given k, let $d_k = \min\{l : V_{kl} \neq \emptyset\}$ and $u_k = \max\{l : V_{kl} \neq \emptyset\}$. Then $A_k = \cup_{l=d_k}^{u_k} V_{kl}$. Moreover, for any $d_k < l < u_k$,

$$\beta^0_{A,k} \leq \max_{i\in B_{d_k}} \beta^0_i \leq \min_{j\in B_l} \beta^0_j \leq \max_{j\in B_l} \beta^0_j \leq \min_{i\in B_{u_k}} \beta^0_i \leq \beta^0_{A,k},$$

where the first and last inequalities are because $A_k \cap B_{d_k} \neq \emptyset$ and $A_k \cap B_{u_k} \neq \emptyset$, and the inequalities in between come from Definition 2.3. It follows that $\beta^0_j = \beta^0_{A,k}$ for all $j \in B_l$. This means $B_l \subset A_k$, and hence $V_{kl} = B_l$.

Consider the second claim. Given l, let $a_l = \min\{k : V_{kl} \neq \emptyset\}$ and $b_l = \max\{k : V_{kl} \neq \emptyset\}$, and hence, $B_l = \cup_{k=a_l}^{b_l} V_{kl}$. For any $a_l < k < b_l$ and $l' < l$,

$$\max_{i \in B_{l'}} \beta_i^0 \leq \min_{i \in B_l} \beta_i^0 \leq \beta_{A,a_l}^0 < \beta_{A,k}^0,$$

where the first inequality comes from Definition 2.3, the second inequality is because $A_{a_l} \cap B_l \neq \emptyset$ and the last inequality is from $\beta_{A,1}^0 < \beta_{A,2}^0 < \cdots < \beta_{A,K}^0$ and $a_l < k$. It follows that $B_{l'} \cap A_k = \emptyset$. Similarly, for any $l' > l$, $B_{l'} \cap A_k = \emptyset$. As a result, $A_k \subset B_l$ and $V_{kl} = A_k$. □

B.3 Proof of Theorem 4.1

Recall the mappings T, T^{-1} and T^* defined in the proof of Theorem 3.1. Write $Q_n(\boldsymbol{\beta}) = L_n(\boldsymbol{\beta}) + P_n(\boldsymbol{\beta})$, where $L_n(\boldsymbol{\beta}) = \frac{1}{2n}\|\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\|^2$ and $P_n(\boldsymbol{\beta}) = P_{\Upsilon,\lambda_1,\lambda_2}(\boldsymbol{\beta})$. For any $\boldsymbol{\mu} \in \mathbb{R}^K$, let

$$L_n^A(\boldsymbol{\mu}) = L_n(T^{-1}(\boldsymbol{\mu})), \qquad P_n^A(\boldsymbol{\mu}) = P_n(T^{-1}(\boldsymbol{\mu})),$$

and define $Q_n^A(\boldsymbol{\mu}) = L_n^A(\boldsymbol{\mu}) + P_n^A(\boldsymbol{\mu})$.

We only need to show that $\widehat{\boldsymbol{\beta}}^{oracle}$ is a strictly local minimizer of Q_n with probability at least $1 - \epsilon_0 - n^{-1}K - 2(n \vee p)^{-1}$. Let E_1' be the event that the segmentation Υ preserves the order of $\boldsymbol{\beta}^0$, and define the event E_2 and $\mathcal{B}$, a neighborhood of $\boldsymbol{\beta}^0$, the same as in the proof of Theorem 3.1. Recall the statements (a) and (b) in the proof of Theorem 3.1. For an event E_3' to be defined such that $P((E_3')^c) \leq 2(n \vee p)^{-1}$, we shall show that (a) and (b) hold on the event $E_1' \cap E_2 \cap E_3'$. The conclusion then follows immediately.

Consider (a) first. Same as before, it suffices to show (29). Recall that $V_{kl} = A_k \cap B_l$. Define $m_{1,kk'} = \sum_{l=1}^{L-1}(|V_{kl}||V_{k'(l+1)}| + |V_{k'l}||V_{k(l+1)}|)$ and $m_{2,kk'} = \sum_{l=1}^{L}|V_{kl}||V_{k'l}|$, for $1 \leq k < k' \leq K$. Write for short $\rho_1(\cdot) = \rho_{\lambda_1}(\cdot)$ and $\rho_2(\cdot) = \rho_{\lambda_2}(\cdot)$. It follows that

$$P_n^A(\boldsymbol{\mu}) = \lambda_1 \sum_{1 \leq k < k' \leq K} m_{1,kk'}\rho_1(|\mu_k - \mu_{k'}|) + \lambda_2 \sum_{1 \leq k < k' \leq K} m_{2,kk'}\rho_2(|\mu_k - \mu_{k'}|).$$

Therefore, it suffices to check

$$\min_{k \neq k'} |\mu_k - \mu_{k'}| > a \max\{\lambda_{1n}, \lambda_{2n}\}, \qquad \text{for any } \boldsymbol{\beta} \in \mathcal{B}, \boldsymbol{\mu} = T^*(\boldsymbol{\beta}).$$

The left hand side is lower bounded by $2b_n - \|\boldsymbol{\beta} - \boldsymbol{\beta}^0\|_\infty \geq 2b_n - C\sqrt{K\log(n)/n} \gg b_n > a\max\{\lambda_{1n}, \lambda_{2n}\}$, which proves (29).

Next, we consider (b). Same as before, it suffices to show (33). For $\boldsymbol{\beta} \in \mathcal{B}$, denote by $\boldsymbol{\beta}^* = T^{-1} \circ T^*(\boldsymbol{\mu})$ its orthogonal projection onto $\mathcal{M}_A$. By Taylor expansion,

$$\begin{aligned} Q_n(\boldsymbol{\beta}) - Q_n(\boldsymbol{\beta}^*) &= -\frac{1}{n}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}^m)^T \mathbf{X}(\boldsymbol{\beta} - \boldsymbol{\beta}^*) + \sum_{j=1}^{p} \frac{\partial P_n(\boldsymbol{\beta}^m)}{\partial \beta_j}(\beta_j - \beta_j^*) \\ &\equiv K_1 + K_2, \end{aligned}$$

where $\boldsymbol{\beta}^m$ is in the line between $\boldsymbol{\beta}$ and $\boldsymbol{\beta}^*$. Let $\bar{\rho}_i(t) = \rho_i'(|t|)\text{sgn}(t)$, $i = 1, 2$. Rearranging the sums in K_2, we can write

$$\begin{aligned} K_2 &= \lambda_1 \sum_{l=1}^{L-1} \sum_{i \in B_l, j \in B_{l+1}} \bar{\rho}_1(\beta_i^m - \beta_j^m)\big[(\beta_i - \beta_j) - (\beta_i^* - \beta_j^*)\big] \\ &\quad + \lambda_2 \sum_{l=1}^{L} \sum_{i,j \in B_l} \bar{\rho}_2(\beta_i^m - \beta_j^m)\big[(\beta_i - \beta_j) - (\beta_i^* - \beta_j^*)\big]. \end{aligned}$$

For those (i, j) not belonging to the same true group, $|\beta_i^m - \beta_j^m| \geq 2b_n - 2\|\boldsymbol{\beta}^m - \boldsymbol{\beta}^0\|_\infty \geq 2b_n - 2\|\boldsymbol{\beta}^* - \boldsymbol{\beta}^0\|_\infty \geq 2b_n - 2\|\boldsymbol{\beta}^* - \boldsymbol{\beta}^0\| \geq 2b_n - 2\|\boldsymbol{\beta} - \boldsymbol{\beta}^0\| > 2b_n - C\sqrt{K \log(n)/n}$. From the conditions on $(b_n, \lambda_{1n}, \lambda_{2n})$, it is easy to see that $\rho_l(|\beta_i^m - \beta_j^m|) = 0$, $l = 1, 2$. On the other hand, for those (i, j) belonging to the same true group, $\beta_i^* = \beta_j^*$ and hence $\text{sgn}(\beta_i^m - \beta_j^m) = \text{sgn}(\beta_i - \beta_j)$. Together, we find that

$$\begin{aligned} K_2 &= \lambda_1 \sum_{l=1}^{L-1} \sum_{i \in B_l, j \in B_{l+1}, i \overset{\mathcal{A}}{\sim} j} \rho_1'(|\beta_i^m - \beta_j^m|)|\beta_i - \beta_j| + \lambda_2 \sum_{l=1}^{L} \sum_{i,j \in B_l, i \overset{\mathcal{A}}{\sim} j} \rho_2'(|\beta_i^m - \beta_j^m|)|\beta_i - \beta_j| \\ &\geq \lambda_1 \sum_{l=1}^{L-1} \sum_{i \in B_l, j \in B_{l+1}, i \overset{\mathcal{A}}{\sim} j} \rho_1'(2t_n)|\beta_i - \beta_j| + \lambda_2 \sum_{l=1}^{L} \sum_{i,j \in B_l, i \overset{\mathcal{A}}{\sim} j} \rho_2'(2t_n)|\beta_i - \beta_j|, \qquad (4) \end{aligned}$$

where $i \overset{\mathcal{A}}{\sim} j$ means i and j are in the same true group, and the last inequality comes from the concavity of ρ and the fact that $|\beta_i^m - \beta_j^m| \leq 2\|\boldsymbol{\beta}^m - \boldsymbol{\beta}^*\|_\infty \leq 2\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_\infty \leq 2t_n$.

Now, we simplify K_1. Let $\mathbf{z} = \mathbf{z}(\boldsymbol{\beta}^m) = \mathbf{X}^T(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}^m)$ and write $K_1 = -\frac{1}{n}\mathbf{z}^T(\boldsymbol{\beta} - \boldsymbol{\beta}^*)$. Note that for each $j \in A_k$, $\beta_j^* = \frac{1}{|A_k|}\sum_{i \in A_k} \beta_i = \frac{1}{|A_k|}\sum_{l=d_k}^{u_k} \sum_{i \in V_{kl}} \beta_i$, where V_{kl}, d_k and u_k are as in Proposition B.1.

$$\begin{aligned} K_1 &= -\frac{1}{n}\sum_{k=1}^{K}\sum_{l=d_k}^{u_k}\sum_{j \in V_{kl}} z_j(\beta_j - \beta_j^*) \\ &= -\frac{1}{n}\sum_{k=1}^{K}\sum_{l=d_k}^{u_k}\sum_{j \in V_{kl}} z_j \frac{1}{|A_k|}\sum_{l'=d_k}^{u_k}\sum_{j' \in V_{kl'}} (\beta_j - \beta_{j'}) \\ &= -\frac{1}{2n}\sum_{k=1}^{K}\frac{1}{|A_k|}\sum_{l=d_k}^{u_k}\sum_{l'=d_k}^{u_k}\sum_{j \in V_{kl}}\sum_{j' \in V_{kl'}} (z_j - z_{j'})(\beta_j - \beta_{j'}) \end{aligned}$$

$$
\begin{aligned}
&= -\frac{1}{n}\sum_{k=1}^{K}\frac{1}{|A_k|}\sum_{l=d_k}^{u_k}\sum_{j,j'\in V_{kl}}(z_j-z_{j'})(\beta_j-\beta_{j'}) \\
&\quad -\frac{1}{n}\sum_{k=1}^{K}\frac{1}{|A_k|}\sum_{d_k\le l<l'\le u_k}\sum_{j\in V_{kl},j'\in V_{kl'}}(z_j-z_{j'})(\beta_j-\beta_{j'}) \\
&\equiv K_{11}+K_{12}.
\end{aligned}
$$

Using notations in Proposition B.1, $\sum_{k=1}^{K}\sum_{l=d_k}^{u_k}=\sum_{l=1}^{L}\sum_{k=a_l}^{b_l}$. Therefore,

$$
\begin{aligned}
K_{11} &= -\frac{1}{n}\sum_{l=1}^{L}\sum_{k=a_l}^{b_l}\sum_{j,j'\in V_{kl}}\frac{1}{|A_k|}(z_j-z_{j'})(\mu_j-\mu_{j'}) \\
&= -\frac{1}{n}\sum_{l=1}^{L}\sum_{j,j'\in B_l,j\overset{\mathcal{A}}{\sim}j'}\theta_{jj'}(\mathbf{z})(\mu_j-\mu_{j'}), \qquad (5)
\end{aligned}
$$

where $\theta_{jj'}(\mathbf{z})\equiv\frac{1}{|A_k|}(z_j-z_{j'})$ for $j,j'\in A_k$. To simplify K_{12}, note that given any (j,j') such that $j\in V_{kl}$ and $j'\in V_{kl'}$, for some k and $l<l'$, we have

$$
\beta_j-\beta_{j'}=\frac{1}{\prod_{h=l+1}^{l'-1}|V_{kh}|}\sum_{\left\{\substack{(i_l,i_{l+1},\cdots,i_{l'}):\ i_l=j,\ i_{l'}=j';\\ i_h\in V_{kh},h=l+1,\cdots,l'-1}\right\}}\sum_{h=l}^{l'-1}(\beta_{i_h}-\beta_{i_{h+1}}).
$$

Plugging this into the expression K_{12}, we obtain

$$
\begin{aligned}
K_{12} &= -\frac{1}{n}\sum_{k=1}^{K}\frac{1}{|A_k|}\sum_{d_k\le l<l'\le u_k}\sum_{\{(i_l,i_{l+1},\cdots,i_{l'}):\ i_h\in V_{kh}\}}\frac{(z_{i_l}-z_{i_{l'}})}{\prod_{h=l+1}^{l'-1}|V_{kh}|}\sum_{h=l}^{l'-1}(\beta_{i_h}-\beta_{i_{h+1}}) \\
&= -\frac{1}{n}\sum_{k=1}^{K}\frac{1}{|A_k|}\sum_{d_k\le l<l'\le u_k}\sum_{h=l}^{l'-1}\sum_{j\in V_{kh},j'\in V_{k(h+1)}}\omega_{jj',ll'h}(\mathbf{z})(\beta_j-\beta_{j'}),
\end{aligned}
$$

where for (j,j',l,l',h) such that $j\in V_{kh}$, $j'\in V_{k(h+1)}$ and $l\le h\le l'-1$,

$$
\omega_{jj',ll'h}(\mathbf{z})=\begin{cases}
z_j-z_{j'}, & l=h=l'-1\\
\frac{|V_{kl'}|}{|V_{k(l+1)}|}(z_j-\bar{z}_{kl'}), & l=h<l'-1\\
\frac{|V_{kl}||V_{kl'}|}{|V_{kh}||V_{k(h+1)}|}(\bar{z}_{kl}-\bar{z}_{kl'}), & l<h<l'-1\\
\frac{|V_{kl}|}{|V_{k(l'-1)}|}(\bar{z}_{kl}-z_{j'}), & l<h=l'-1
\end{cases},
$$

and $\bar{z}_{kl}$ is the average of $\{z_j:j\in V_{kl}\}$. By rearranging terms, $\sum_{k=1}^{K}\sum_{d_k\le l<l'\le u_k}\sum_{h=l}^{l'-1}=\sum_{h=1}^{L-1}\sum_{k=a_h}^{b_h}\sum_{(l,l'):d_k\le l\le h<l'\le u_k}$. Therefore,

$$
K_{12} = -\frac{1}{n}\sum_{h=1}^{L-1}\sum_{k=a_h}^{b_h}\frac{1}{|A_k|}\sum_{j\in V_{kh},j'\in V_{k(h+1)}}\Big[\sum_{l=d_k}^{h}\sum_{l'=h+1}^{u_k}\omega_{jj',ll'h}(\mathbf{z})\Big](\beta_j-\beta_{j'})
$$

$$= -\frac{1}{n}\sum_{h=1}^{L-1}\sum_{j\in B_h, j'\in B_{h+1}, j\overset{\mathcal{A}}{\sim}j'} \tau_{jj'}(\mathbf{z})(\beta_j - \beta_{j'}), \tag{6}$$

where

$$\begin{aligned}
\tau_{jj'}(\mathbf{z}) &= \frac{1}{|A_k|}\sum_{l=d_k}^{h}\sum_{l'=h+1}^{u_k} \omega_{jj',ll'h}(\mathbf{z}) \\
&= \frac{1}{|A_k|}\sum_{l=d_k}^{h-1}\sum_{l'=h+2}^{u_k} \frac{|V_{kl}||V_{kl'}|}{|V_{kh}||V_{k(h+1)}|}(\bar{z}_{kl} - \bar{z}_{kl'}) + \frac{1}{|A_k|}\sum_{l=d_k}^{h-1}\frac{|V_{kl}|}{|V_{kh}|}(\bar{z}_{kl} - z_{j'}) \\
&\quad + \frac{1}{|A_k|}\sum_{l'=h+2}^{u_k}\frac{|V_{kl'}|}{|V_{k(h+1)}|}(z_j - \bar{z}_{kl'}) + \frac{1}{|A_k|}(z_j - z_{j'}) \\
&= \frac{1}{|A_k|}\sum_{l=d_k}^{h-1}\frac{|V_{kl}|(\sum_{l'=h+1}^{u_k}|V_{kl'}|)}{|V_{kh}||V_{k(h+1)}|}\bar{z}_{kl} + \frac{1}{|A_k|}\frac{\sum_{l'=h+1}^{u_k}|V_{kl'}|}{|V_{k(h+1)}|}z_j \\
&\quad - \frac{1}{|A_k|}\sum_{l'=h+2}^{u_k}\frac{(\sum_{l=d_k}^{h}|V_{kl}|)|V_{kl'}|}{|V_{kh}||V_{k(h+1)}|}\bar{z}_{kl'} - \frac{1}{|A_k|}\frac{\sum_{l=d_k}^{h}|V_{kl}|}{|V_{kh}|}z_{j'}.
\end{aligned}$$

Let $A_{kh}^1 = \cup_{l\le h} V_{kl}$ and $A_{kh}^2 = \cup_{l>h} V_{kl}$. Then, for any (j, j') such that $j \in B_h$, $j' \in B_{h+1}$ and $j, j' \in A_k$, we have the following expression

$$\begin{aligned}
\tau_{jj'}(\mathbf{z}) &= \frac{1}{|V_{kh}||V_{k(h+1)}|}\Big(\frac{|A_{kh}^2|}{|A_k|}\sum_{i\in A_{k(h-1)}^1} z_i - \frac{|A_{kh}^1|}{|A_k|}\sum_{i\in A_{k(h+1)}^2} z_i\Big) \\
&\quad + \Big(\frac{|A_{kh}^2|}{|A_k||V_{k(h+1)}|}z_j - \frac{|A_{kh}^1|}{|A_k||V_{kh}|}z_{j'}\Big).
\end{aligned} \tag{7}$$

Combining (5) and (6) gives

$$|K_1| \le \frac{1}{n}\sum_{l=1}^{L-1}\sum_{\substack{i\in B_l, j\in B_{l+1},\\ i\overset{\mathcal{A}}{\sim}j}} |\tau_{ij}(\mathbf{z})||\beta_i - \beta_j| + \frac{1}{n}\sum_{l=1}^{L}\sum_{i,j\in B_l, i\overset{\mathcal{A}}{\sim}j} |\theta_{ij}(\mathbf{z})||\beta_i - \beta_j|. \tag{8}$$

Using the inequalities on K_1 and K_2, i.e., (4) and (8), we have

$$\begin{aligned}
Q_n(\boldsymbol{\beta}) - Q_n(\boldsymbol{\beta}^*) &\ge \sum_{l=1}^{L-1}\sum_{i\in B_l, j\in B_{l+1}, i\overset{\mathcal{A}}{\sim}j} \big[\lambda_1\rho_1'(2t_n) - n^{-1}\tau_{ij}(\mathbf{z})\big]|\beta_i - \beta_j| \\
&\quad + \sum_{l=1}^{L}\sum_{i,j\in B_l, i\overset{\mathcal{A}}{\sim}j} \big[\lambda_2\rho_2'(2t_n) - n^{-1}\theta_{ij}(\mathbf{z})\big]|\beta_i - \beta_j|.
\end{aligned}$$

Therefore, showing (33) reduces to showing that, over the event $E_1' \cap E_2$, for sufficiently small t_n,

$$n^{-1}\max_{ij}|\tau_{ij}(\mathbf{z})| < \lambda_1\rho_1'(2t_n) \quad \text{and} \quad n^{-1}\max_{ij}|\theta_{ij}(\mathbf{z})| < \lambda_2\rho_2'(2t_n), \tag{9}$$

except for a probability of at most $2(n \vee p)^{-1}$.

Note that $\mathbf{z} = \mathbf{X}^T\boldsymbol{\varepsilon} - \boldsymbol{\eta} - \boldsymbol{\eta}^m$, where $\boldsymbol{\eta} = \mathbf{X}^T\mathbf{X}(\boldsymbol{\beta}^* - \boldsymbol{\beta}^0)$ and $\boldsymbol{\eta}^m = \mathbf{X}^T\mathbf{X}(\boldsymbol{\beta}^m - \boldsymbol{\beta}^*)$. It is seen that $\|\boldsymbol{\eta}^m\| \leq \lambda_{\max}(\mathbf{X}^T\mathbf{X})\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\| \leq \lambda_{\max}(\mathbf{X}^T\mathbf{X})t_n$. So $\tau_{ij}(\mathbf{z}) = \tau_{ij}(\mathbf{X}^T\boldsymbol{\varepsilon} + \boldsymbol{\eta}) + rem$, where the remainder term is uniformly bounded by $g_n(t_n)$, for some function $g_n(\cdot)$ such that $g_n(0+) = 0$. Similar situations are observed for $\theta_{ij}(\mathbf{z})$. As a result, to show (9), it suffices to show that over the event $E_1' \cap E_2$,

$$n^{-1} \max_{ij} |\theta_{ij}(\mathbf{X}^T\boldsymbol{\varepsilon} + \boldsymbol{\eta})| < \lambda_2 \rho_2'(0+), \tag{10}$$

and

$$n^{-1} \max_{ij} |\tau_{ij}(\mathbf{X}^T\boldsymbol{\varepsilon} + \boldsymbol{\eta})| < \lambda_1 \rho_1'(0+), \tag{11}$$

except for a probability of at most $2(n \vee p)^{-1}$.

First, consider (10). Let E_{31}' be the event

$$n^{-1} \max_{i,j \in A_k} |\theta_{ij}(\mathbf{X}^T\boldsymbol{\varepsilon})| \leq |A_k|^{-1}\sqrt{6c_3^{-1}\log(2(n \vee p))/n}, \quad \text{for all } k.$$

Note that $\theta_{ij}(\mathbf{X}^T\boldsymbol{\varepsilon}) = \frac{1}{|A_k|}(\mathbf{x}_i - \mathbf{x}_j)^T\boldsymbol{\varepsilon}$, where $\|\mathbf{x}_i - \mathbf{x}_j\| \leq \sqrt{2n}$. Applying Condition 3.3 and the union bound,

$$P((E_{31}')^c) \leq \sum_{k=1}^{K} \sum_{i,j \in A_k} P\left((\mathbf{x}_i - \mathbf{x}_j)^T\boldsymbol{\varepsilon} > \|\mathbf{x}_i - \mathbf{x}_j\|\sqrt{3c_3^{-1}\log(2(n \vee p))}\right) < (n \vee p)^{-1}.$$

Moreover, $|\theta_{ij}(\boldsymbol{\eta})| \leq \frac{2}{|A_k|}\max_{i'} |\eta_{i'} - \bar{\eta}_k|$, where $\bar{\eta}_k$ is the average of $\{\eta_i : i \in A_k\}$. Note that $\max_{i \in A_k} |\eta_i - \bar{\eta}_k| \leq n\nu_k\|\boldsymbol{\beta}^* - \boldsymbol{\beta}^0\|$ and $\|\boldsymbol{\beta}^* - \boldsymbol{\beta}^0\| \leq \|\boldsymbol{\beta} - \boldsymbol{\beta}^0\|$ because $\boldsymbol{\beta}^*$ is the orthogonal projection of $\boldsymbol{\beta}$ onto $\mathcal{M}_A$. Noticing that $\boldsymbol{\beta} \in \mathcal{B}$, we obtain

$$n^{-1} \max_{i,j} |\theta_{ij}(\boldsymbol{\eta})| \leq C\nu_k |A_k|^{-1}\sqrt{K\log(n)/n}.$$

Combing the above results to the choice of λ_2 gives $n^{-1}\max_{i,j} |\theta_{ij}(\mathbf{z})| \ll \lambda_2$, and (10) follows.

Next, consider (11). First, we bound $\tau_{jj'}(\mathbf{X}^T\boldsymbol{\varepsilon})$. From (7), $\tau_{jj'}(\mathbf{X}^T\boldsymbol{\varepsilon}) = \tilde{\mathbf{a}}_{jj'}^T\boldsymbol{\varepsilon}$, where

$$\begin{aligned}\tilde{\mathbf{a}}_{jj'} &= \frac{1}{|V_{kh}||V_{k(h+1)}|}\left[\frac{|A_{kh}^2|}{|A_k|}\mathbf{X}_{A_{k(h-1)}^1}\mathbf{1}_{A_{k(h-1)}^1} - \frac{|A_{kh}^1|}{|A_k|}\mathbf{X}_{A_{k(h+1)}^2}\mathbf{1}_{A_{k(h+1)}^2}\right] \\ &\quad + \frac{|A_{kh}^2|}{|A_k||V_{k(h+1)}|}\mathbf{x}_j - \frac{|A_{kh}^1|}{|A_k||V_{kh}|}\mathbf{x}_{j'}.\end{aligned}$$

Recall that $n\sigma_k$ is the maximum eigenvalue of $\mathbf{X}_{A_k}^T\mathbf{X}_{A_k}$. It follows that

$$\|\tilde{\mathbf{a}}_{jj'}\|^2 \leq 4n\sigma_k\left(\frac{|A_{kh}^2|^2|A_{k(h-1)}^1| + |A_{kh}^1|^2|A_{k(h+1)}^2|}{|V_{kh}|^2|V_{k(h+1)}|^2|A_k|^2} + \frac{|A_{kh}^2|^2}{|A_k|^2|V_{k(h+1)}|^2} + \frac{|A_{kh}^1|^2}{|A_k|^2|V_{kh}|^2}\right)$$

$$
\begin{aligned}
&\leq \quad 4n\sigma_k \begin{cases} \frac{|A_k|}{|B_h|^2|B_{h+1}|^2} + \frac{1}{|B_{h+1}|^2} + \frac{1}{|B_h|^2}, & h > d_k,\ h+1 < u_k \\ \frac{|A_k|}{|B_h|^2|V_{k(h+1)}|^2} + \frac{1}{|A_k|^2} + \frac{1}{|B_h|^2}, & h > d_k,\ h+1 = u_k \\ \frac{|A_k|}{|V_{kh}|^2|B_{h+1}|^2} + \frac{1}{|B_{h+1}|^2} + \frac{1}{|A_k|^2}, & h = d_k,\ h+1 < u_k \\ \frac{2}{|A_k|^2}, & h = d_k,\ h+1 = u_k \end{cases} \\
&\leq \quad n\sigma_k \frac{12|A_k|}{\min\{|A_k|^3,\ \min_{d_k \leq h \leq u_k}\{|B_h|^2\}\}} \quad = \quad 12n\sigma_k\phi_k. \qquad (12)
\end{aligned}
$$

Here in the second inequality, we have used the following facts: (1)From Proposition B.1, for $d_k < h < u_k$, $|V_{kh}| = |B_h|$ and $|V_{k(h+1)}| = |B_{h+1}|$. (2) When $h = d_k$, $|A^1_{kh}| = |V_{kh}|$; when $h+1 = u_k$, $|A^2_{kh}| = |V_{k(h+1)}|$. (3) $|A^1_{k(h-1)}| < |A^1_{kh}| \leq |A_k|$, $|A^2_{k(h+1)}| < |A^2_{kh}| \leq |A_k|$, and $|A^1_{kh}| + |A^2_{kh}| = |A_k|$. In the third inequality, we have used the fact that $|V_{kh}| \geq 1$ when $V_{kh} \neq \emptyset$. Let E'_{32} be the event that

$$
n^{-1} \max_{j,j'} |\tau_{jj'}(\mathbf{X}^T\boldsymbol{\varepsilon})| \leq C\sqrt{\sigma_k\phi_k \log(n \vee p)/n}, \qquad \text{for all } k. \qquad (13)
$$

Applying Condition 3.3, (12) and the union bound, it is easy to see that $P((E'_{32})^c) < (n \vee p)^{-1}$ for some large enough constant $C > 0$.

Second, we bound $\tau_{jj'}(\boldsymbol{\eta})$. We observe from (7) that $\tau_{jj'}(\mathbf{v}) = 0$, for any $\mathbf{v}$ with equal elements in A_k. Thus, $\tau_{jj'}(\boldsymbol{\eta}) = \tau_{jj'}(\boldsymbol{\eta} - \bar{\eta}_k\mathbf{1})$, where $\bar{\eta}_k$ is the average over the elements of $\boldsymbol{\eta}$ in A_k. By similarly analysis to that in (12), we find that

$$
|\tau_{jj'}(\boldsymbol{\eta})|^2 = |\tau_{jj'}(\boldsymbol{\eta} - \bar{\eta}_k\mathbf{1})|^2 \leq 12\phi_k\big(\max_{i\in A_k}\{|\eta_i - \bar{\eta}_k|\}\big)^2.
$$

By definition, $\max_{i\in A_k}\{|\eta_i - \bar{\eta}_k|\} \leq n\nu_k\|\boldsymbol{\beta}^* - \boldsymbol{\beta}^0\| \leq C\nu_k\sqrt{nK\log(n)}$. It follows that

$$
n^{-1} \max_{j,j'} |\tau_{jj'}(\boldsymbol{\eta})| \leq C\nu_k\sqrt{u_k K \log(n)/n}. \qquad (14)
$$

Combining (13) and(14), we then obtain (11) from the condition on λ_1. □

B.4 Proof of Theorem 4.2

Since $\widehat{\boldsymbol{\beta}}^{oracle}_A - \boldsymbol{\beta}^0 = (\mathbf{X}_A^T\mathbf{X}_A)^{-1}(\mathbf{X}_A^T\boldsymbol{\varepsilon})$, to show the claim, it suffices to show

$$
\mathbf{B}_n(\mathbf{X}_A^T\mathbf{X}_A)^{-1/2}\mathbf{X}_A^T\boldsymbol{\varepsilon} \xrightarrow{d} N(\mathbf{0}, \mathbf{H}).
$$

Equivalently, for any $\mathbf{a} \in \mathbb{R}^q$,

$$
\mathbf{a}^T\mathbf{B}_n(\mathbf{X}_A^T\mathbf{X}_A)^{-1/2}\mathbf{X}_A^T\boldsymbol{\varepsilon} \xrightarrow{d} N(0, \mathbf{a}^T\mathbf{H}\mathbf{a}). \qquad (15)
$$

Let $\mathbf{v} = \mathbf{X}_A(\mathbf{X}_A^T\mathbf{X}_A)^{-1/2}\mathbf{B}_n^T\mathbf{a}$, and write the left hand side of (15) as $\mathbf{v}^T\boldsymbol{\varepsilon} = \sum_{i=1}^n v_i\varepsilon_i$. The $v_i\varepsilon_i$'s are independently distributed with $E[v_i\varepsilon_i] = 0$ and $E[|v_i\varepsilon_i|^2] = v_i^2$. Let $s_n^2 = \sum_{i=1}^n E[|v_i\varepsilon_i|^2]$. By Lindeberg's central limit theorem, if for any $\epsilon > 0$,

$$\lim_{n\to\infty} s_n^{-2}E\big[|v_i\varepsilon_i|^2 1\{|v_i\varepsilon_i| > \epsilon s_n\}\big] = 0, \tag{16}$$

then $s_n^{-1}\sum_{i=1}^n v_i\varepsilon_i \xrightarrow{d} N(0,1)$. Since $s_n^2 = \mathbf{a}^T\mathbf{B}_n\mathbf{B}_n^T\mathbf{a} \to \mathbf{a}^T\mathbf{H}\mathbf{a}$, (15) follows immediately from the Slutsky's lemma.

It remains to show (16). Using the formula $E[X1\{X > \epsilon\}] = \epsilon P(X > \epsilon) + \int_\epsilon^\infty P(X > u)du$ for $X = |v_i\varepsilon_i|^2$, we have

$$E\big[|v_i\varepsilon_i|^2 1\{|v_i\varepsilon_i| > \epsilon s_n\}\big] = \epsilon^2 s_n^2 P(|v_i\varepsilon_i| > \epsilon s_n) + \int_{\epsilon s_n}^\infty P(|v_i\varepsilon_i| > \sqrt{u})du.$$

From Condition 3.3,

$$P(|v_i\varepsilon_i| > \epsilon s_n) \le 2e^{-c_3\epsilon^2 s_n^2/|v_i|^2} \le \frac{2|v_i|^4}{c_3^2\epsilon^4 s_n^4},$$

where the last inequality is due to that $\exp(-x) \le x^{-k}$ for any $x > 0$ and positive integer k. Similarly,

$$\int_{\epsilon s_n}^\infty P(|v_i\varepsilon_i| > \sqrt{u})du \le 2\int_{\epsilon s_n}^\infty e^{-c_3 u/|v_i|^2}du = \frac{2|v_i|^2}{c_3}e^{-c_3\epsilon s_n/|v_i|^2} \le \frac{2|v_2|^4}{c_3\epsilon s_n}.$$

Note that $s_n^{-1} = O(1)$ since $s_n \to \mathbf{a}^T\mathbf{H}\mathbf{a}$. We have

$$\begin{aligned} & \frac{1}{s_n^2}\sum_{i=1}^n E\big[|v_i\varepsilon_i|^2 1\{|v_i\varepsilon_i| > \epsilon s_n\}\big] \\ \le\ & C\sum_{i=1}^n |v_i|^4 \ = \ C\|\mathbf{X}_A(\mathbf{X}_A^T\mathbf{X}_A)^{-1/2}\mathbf{B}_n^T\|_4^4 \\ \le\ & C\big(\|\mathbf{X}_A(\mathbf{X}_A^T\mathbf{X}_A)^{-1/2}\mathbf{B}_n^T\|_{2,4}\cdot\|\mathbf{a}\|\big)^4. \end{aligned}$$

The right hand side is $o(1)$ by assumption. This proves (16). □

B.5 Proof of Corollary 4.1

It is easy to see that the asymptotic variance of $\mathbf{a}_n^T(\widehat{\boldsymbol{\beta}}^{ols} - \boldsymbol{\beta}^0)$ is $\mathbf{a}_n^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{a}_n = v_{1n}$. Consider $\mathbf{a}_n^T(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)$. Noting that $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0 = \mathbf{M}_n\mathbf{D}(\widehat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_A^0)$, we can write

$$\mathbf{a}_n^T(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) = \mathbf{a}_n^T\mathbf{M}_n\mathbf{D}(\mathbf{X}_A^T\mathbf{X}_A)^{-1/2}(\mathbf{X}_A^T\mathbf{X}_A)^{1/2}(\widehat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_A^0),$$

where $\mathbf{D} = \mathrm{diag}(|A_1|^{1/2}, \cdots, |A_K|^{1/2})$. Take $\mathbf{B}_n = \mathbf{a}_n^T\mathbf{M}_n\mathbf{D}(\mathbf{X}_A^T\mathbf{X}_A)^{-1/2}$ and apply Theorem 4.2. It implies that the asymptotic variance of $\mathbf{a}_n^T(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)$ is

$$\mathbf{B}_n\mathbf{B}_n^T = \mathbf{a}_n^T\mathbf{M}_n\mathbf{D}(\mathbf{X}_A^T\mathbf{X}_A)^{-1}\mathbf{D}\mathbf{M}_n^T\mathbf{a}_n.$$

Observing that $\mathbf{X}_A = \mathbf{X}\mathbf{M}_n\mathbf{D}$, the above quantity is equal to

$$\mathbf{a}_n^T \mathbf{M}_n (\mathbf{M}_n^T \mathbf{X}^T \mathbf{X} \mathbf{M}_n)^{-1} \mathbf{M}_n^T \mathbf{a}_n = v_{2n}.$$

Next, we show $v_{1n} > v_{2n}$. Since $\mathbf{M}_n^T \mathbf{M}_n = \mathbf{I}_K$, there exists an orthogonal matrix $\mathbf{Q}$ such that $\mathbf{M}_n$ is equal to the first K columns of $\mathbf{Q}$. Write $\mathbf{b} = \mathbf{Q}^T \mathbf{a}_n$ and $\mathbf{G} = \mathbf{Q}^T \mathbf{X}^T \mathbf{X} \mathbf{Q}$. Direct calculations yield $v_{1n} = \mathbf{b}^T \mathbf{G}^{-1} \mathbf{b}$ and $v_{2n} = \mathbf{b}_1^T \mathbf{G}_{11}^{-1} \mathbf{b}_1$, where $\mathbf{b}_1$ is the subvector of $\mathbf{v}$ formed by its first K elements and $\mathbf{G}_{11}$ is the upper left $K \times K$ block of $\mathbf{G}$. From basic algebra, $v_{1n} \geq v_{2n}$. □

B.6 Proof of Theorem 4.3

The proof of $\|\widehat{\boldsymbol{\beta}}^{oracle} - \boldsymbol{\beta}^0\| = O_p(\sqrt{K/n})$ is the same as that in Theorem 3.1. We only need to show that $\widehat{\boldsymbol{\beta}}^{oracle}$ is a strictly local minimizer of Q_n^{sparse}, with probability at least $1 - \epsilon_0 - n^{-1}K - (n \vee s)^{-1} - (n \vee \tilde{s})^{-1}$. Without loss of generality, we assume $\widetilde{S} = \{1, \cdots, p\}$ and $\tilde{s} = p$.

Let $\mathcal{B} = \{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\beta}^0\| \leq C\sqrt{K \log(n)/n}\}$, for a sufficiently large constant $C > 0$. By assumption and (25), $\widehat{\boldsymbol{\beta}}^{oracle} \in \mathcal{B}$ except for a probability of at most $(\epsilon_0 + n^{-1}K)$. For any $\boldsymbol{\beta} \in \mathcal{B}$, let $\boldsymbol{\beta}_S$ be the vector such that $\beta_{S,j} = \beta_j 1\{j \in S\}$, where S is the support of $\boldsymbol{\beta}^0$; and let $\boldsymbol{\beta}_S^*$ be the orthogonal projection of $\boldsymbol{\beta}_S$ onto $\mathcal{M}_A^*$, namely, $\beta_{S,j}^* = \frac{1}{|A_k|}\sum_{i \in A_k} \beta_j$ for any $j \in A_k$, and $\beta_{S,j}^* = 0$ for any $j \notin S$. We aim to show that except for a probability of at most $(n \vee s)^{-1} + (n \vee p)^{-1}$,

(a) For any $\boldsymbol{\beta} \in \mathcal{B}$,

$$Q_n^{sparse}(\boldsymbol{\beta}_S^*) \geq Q_n^{sparse}(\widehat{\boldsymbol{\beta}}^{oracle}), \tag{17}$$

and the inequality is strict whenever $\boldsymbol{\beta}_S^* \neq \widehat{\boldsymbol{\beta}}^{oracle}$.

(b) There exists a positive sequence $\{t_n\}$ such that, for any $\boldsymbol{\beta} \in \mathcal{B}$, $\|\boldsymbol{\beta}_S - \widehat{\boldsymbol{\beta}}^{oracle}\| \leq t_n$,

$$Q_n^{sparse}(\boldsymbol{\beta}_S) \geq Q_n^{sparse}(\boldsymbol{\beta}_S^*), \tag{18}$$

and the inequality is strict whenever $\boldsymbol{\beta}_S \neq \boldsymbol{\beta}_S^*$.

(c) There exists a positive sequence $\{t_n'\}$ such that, for any $\boldsymbol{\beta} \in \mathcal{B}$, $\|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}^{oracle}\| \leq t_n'$,

$$Q_n^{sparse}(\boldsymbol{\beta}) \geq Q_n^{sparse}(\boldsymbol{\beta}_S), \tag{19}$$

and the inequality is strict whenever $\boldsymbol{\beta} \neq \boldsymbol{\beta}_S$.

Suppose (a)-(c) hold. Consider the neighborhood of $\widehat{\boldsymbol{\beta}}^{oracle}$ defined as $\mathcal{B}_n = \{\boldsymbol{\beta} \in \mathcal{B} : \|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}^{oracle}\| \leq \min\{t_n, t'_n\}\}$. It is easy to see that $\|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}^{oracle}\| \leq t'_n$ and $\|\boldsymbol{\beta}_S - \widehat{\boldsymbol{\beta}}^{oracle}\| \leq \|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}^{oracle}\| \leq t_n$ for any $\boldsymbol{\beta} \in \mathcal{B}_n$. As a result, $Q_n^{sparse}(\boldsymbol{\beta}) \geq Q_n^{sparse}(\widehat{\boldsymbol{\beta}}^{oracle})$ for $\boldsymbol{\beta} \in \mathcal{B}_n$, and the inequality is strict except that $\boldsymbol{\beta} = \boldsymbol{\beta}_S = \boldsymbol{\beta}^*_S = \widehat{\boldsymbol{\beta}}^{oracle}$. It follows that $\widehat{\boldsymbol{\beta}}^{oracle}$ is a strictly local minimizer of Q_n^{sparse}.

Now, we show (a)-(c). We claim that (a) and (b) hold except for a probability of at most $(n \vee s)^{-1}$. The proofs are exactly the same as those for (27) and (28), by noting that $Q_n^{sparse}(\boldsymbol{\beta}) = Q_n(\boldsymbol{\beta})$ for any $\boldsymbol{\beta} \in \mathcal{B}$ whose support is contained in S. To show (c), note that $\|\boldsymbol{\beta} - \boldsymbol{\beta}_S\| \leq \|\boldsymbol{\beta}_S - \widehat{\boldsymbol{\beta}}^{oracle}\|$, since $\boldsymbol{\beta}_S$ is the projection of $\boldsymbol{\beta}$ onto the coordinate space of S and $\widehat{\boldsymbol{\beta}}^{oracle}$ belongs to this space. So it suffices to show that (19) holds for all $\boldsymbol{\beta} \in \mathcal{B}$ such that $\|\boldsymbol{\beta} - \boldsymbol{\beta}_S\| \leq t'_n$.

By Taylor expansion,

$$Q_n^{sparse}(\boldsymbol{\beta}) - Q_n^{sparse}(\boldsymbol{\beta}_S) = -\frac{1}{n}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}^m)^T \mathbf{X}(\boldsymbol{\beta} - \boldsymbol{\beta}_S) + \lambda_n \sum_{j \notin S} \bar{\rho}(\beta_j^m)\beta_j,$$

where $\boldsymbol{\beta}^m$ lies in the line between $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_S$. Let $\mathbf{z} = \mathbf{z}(\widehat{\boldsymbol{\beta}}^m) = \mathbf{X}^T(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}^m)$. First, note that $\operatorname{sgn}(\beta_j^m) = \operatorname{sgn}(\beta_j)$ for $j \notin S$. Second, $\|\boldsymbol{\beta}^m - \boldsymbol{\beta}_S\| \leq \|\boldsymbol{\beta} - \boldsymbol{\beta}_S\| \leq t'_n$. Hence, for $j \notin S$, $|\beta_j^m| \leq t'_n$. By the concavity of ρ, $\rho'(|\beta_j^m|) \geq \rho'(t'_n)$. Combining the above, we get

$$Q_n^{sparse}(\boldsymbol{\beta}) - Q_n^{sparse}(\boldsymbol{\beta}_S) \geq \sum_{j \notin S} [\lambda_n \rho'(t'_n) - n^{-1}|z_j|]|\beta_j|.$$

Write $\mathbf{z} = \mathbf{X}^T\boldsymbol{\varepsilon} + \boldsymbol{\eta} + \boldsymbol{\eta}^m$, where $\boldsymbol{\eta} = \mathbf{X}^T\mathbf{X}(\boldsymbol{\beta}^0 - \boldsymbol{\beta}_S)$ and $\boldsymbol{\eta}^m = \mathbf{X}^T\mathbf{X}(\boldsymbol{\beta}_S - \boldsymbol{\beta}^m)$. Since $\|\boldsymbol{\beta}_S - \boldsymbol{\beta}^m\| \leq \|\boldsymbol{\beta}_S - \boldsymbol{\beta}\| \leq t'_n$, $\|\boldsymbol{\eta}^m\|_\infty \leq \lambda_{\max}(\mathbf{X}^T\mathbf{X})t'_n$. Consequently,

$$Q_n^{sparse}(\boldsymbol{\beta}) - Q_n^{sparse}(\boldsymbol{\beta}_S) \geq \sum_{j \notin S} \left[\lambda_n \rho'(0+) - n^{-1}\|\mathbf{X}^T\boldsymbol{\varepsilon} + \boldsymbol{\eta}\|_\infty - g_n(t'_n)\right]|\beta_j|,$$

where $g_n(t'_n) = \lambda_n[\rho'(0+) - \rho'(t_n)] + n^{-1}\lambda_{\max}(\mathbf{X}^T\mathbf{X})t'_n$ satisfying $g_n(0) = 0$. Therefore, if

$$n^{-1}\|\mathbf{X}^T\boldsymbol{\varepsilon} + \boldsymbol{\eta}\|_\infty < \lambda_n \rho'(0+), \tag{20}$$

then there always exits sufficiently small t'_n such that (19) holds.

It remains to show (20). First, by Condition 3.3 and applying the probability union bound, $\|\mathbf{X}^T\boldsymbol{\varepsilon}\|_\infty \leq \sqrt{(2n/c_3)\log(2(n \vee p))}$, except for a probability of at most $(n \vee p)^{-1}$. Second, $\|\boldsymbol{\eta}\|_\infty \leq \|\mathbf{X}^T\mathbf{X}_S\|_{2,\infty}\|\boldsymbol{\beta}_S - \boldsymbol{\beta}^0\| \leq \|\mathbf{X}^T\mathbf{X}_S\|_{2,\infty} \cdot C\sqrt{K\log(n)/n}$. where we have used the fact that $\|\boldsymbol{\beta}^0 - \boldsymbol{\beta}_S\| \leq \|\boldsymbol{\beta} - \boldsymbol{\beta}^0\| \leq C\sqrt{K\log(n)/n}$. Combining the two parts,

$$n^{-1}\|\mathbf{X}^T\boldsymbol{\varepsilon} + \boldsymbol{\eta}\|_\infty \leq C\left(\sqrt{\log(n \vee p)/n} + \|\mathbf{X}^T\mathbf{X}_S\|_{2,\infty}\sqrt{K\log(n)/n}\right) \ll \lambda_n,$$

which proves (20). □