Supplementary material to

DISTRIBUTED TESTING AND ESTIMATION UNDER SPARSE HIGH DIMENSIONAL MODELS

By Heather Battey*† and Jianqing Fan* and Han Liu* and Junwei Lu* and Ziwei Zhu*

APPENDIX A: THE LOW-DIMENSIONAL LINEAR MODEL

As mentioned earlier, the infinity norm bound derived in Lemma 4.1 can be used to do model selection, after which the selected support can be shared across all the local agents. We significantly reduce the dimension of the problem as we only need to refit the data on the selected model. The remaining challenge is to implement the divide and conquer strategy in the low dimensional setting, which is also of independent interest. Here we focus on the linear model, while the generalized linear model is covered in Appendix B.

In this section d still stands for dimension, but in contrast with the rest of this paper in which $d \gg n$, here we consider $d < n$. More specifically, we consider the linear model (3.2) with $d < n$ and i.i.d sub-Gaussian noise $\{\varepsilon_i\}_{i=1}^n$. It is well known that the ordinary least square (OLS) estimator of $\boldsymbol{\beta}^*$ is defined as $\widehat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \boldsymbol{Y}$. In the massive data setting, the communication cost of estimating and inverting covariance matrices is very high (order $O(kd^2)$). However, as pointed out by Chen and Xie (2012), this estimator exactly coincides with the DC estimator,

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{j=1}^{k} X^{(j)T} X^{(j)} \right)^{-1} \sum_{j=1}^{k} X^{(j)T} \boldsymbol{Y}^{(j)}.$$

In this section, we study the DC strategy to approximate $\widehat{\boldsymbol{\beta}}$ with the communication cost only $O(kd)$, which implies that we can only communicate d dimensional vectors.

The OLS estimator based on the subsample $\mathcal{D}_j$ is defined as $\widehat{\boldsymbol{\beta}}(\mathcal{D}_j) = (X^{(j)T} X^{(j)})^{-1} X^{(j)T} \boldsymbol{Y}^{(j)}$. In order to estimate $\boldsymbol{\beta}^*$, a simple and natural idea

*Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08540; Email: {hbattey,jqfan,hanliu,junweil,ziweiz} @princeton.edu; †Department of Mathematics, Imperial College London, London, SW7 2AZ; Email: h.battey@imperial.ac.uk

is to take the average of $\{\widehat{\boldsymbol{\beta}}(\mathcal{D}_j)\}_{j=1}^k$, which we denote by $\overline{\boldsymbol{\beta}}$. The question is whether this estimator preserves the statistical error as $\widehat{\boldsymbol{\beta}}$. The following theorem gives an upper bound of the gap between $\overline{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}}$, and shows that this gap is negligible compared with the statistical error of $\widehat{\boldsymbol{\beta}}$ as long as k is not too large.

Here we give intuitive discussion on the source of efficiency loss. According to proof of Theorem A.1, we have

$$\overline{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}} = \frac{1}{k}\sum_{j=1}^k \left(\left(X^{(j)T}X^{(j)}/n_k\right)^{-1} - (X^TX/n)^{-1}\right) X^{(j)T}\boldsymbol{\varepsilon}^{(j)}/n_k.$$

Since $\{X^{(j)T}\boldsymbol{\varepsilon}^{(j)}/n_k\}_{j=1}^k$ are homogeneous and independent to each other conditional on $\boldsymbol{X}$, the efficiency loss incurred by the DC procedure, i.e., the gap $\overline{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}$, is characterized by the difference between $\big(\frac{1}{k}\sum_{j=1}^k \frac{1}{n_k}X^{(j)T}X^{(j)}\big)^{-1}$ and $\frac{1}{k}\sum_{j=1}^k\big(\frac{1}{n_k}X^{(j)T}X^{(j)}\big)^{-1}$. The rate of $\overline{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}$ is studied in detail in subsequent theorems.

Theorem A.1. Consider the linear model (3.2). Suppose Conditions 3.1 and 3.2 hold and $\{\varepsilon_i\}_{i=1}^n$ are i.i.d sub-Gaussian random variables with $\|\varepsilon_i\|_{\psi_2} \le \sigma_1$. If the number of subsamples satisfies $k = O(nd/(d \vee \log n)^2)$, then for sufficiently large n and d it follows that

$$\|\overline{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}\|_2 = O_{\mathbb{P}}\Big(\frac{\sqrt{k}(d\vee\log n)}{n}\Big), \quad \|\overline{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 = O_{\mathbb{P}}\big(\sqrt{d/n}\big). \tag{A.1}$$

Remark A.2. By taking $k = o\big(nd/(d\vee\log n)^2\big)$, the loss incurred by the divide and conquer procedure, i.e., $\|\overline{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}\|_2$, converges at a faster rate than the statistical error of the full sample estimator $\widehat{\boldsymbol{\beta}}$. In another independent work Rosenblatt and Nadler (2016), the authors also reveal a similar phenomenon under the broad family of generative linear models. They show that when $k = o\big(n/d\big)$, $E\|\overline{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2^2/E\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2^2 \to 1$. In other words, there is no first-order loss by divide and conquer.

We now take a different viewpoint by returning to the high dimensional setting of Section 4.1 ($d \gg n$) and applying Theorem A.1 in the context of a refitting estimator. In this refitting setting, the sparsity s of Lemma 4.1 becomes the dimension of a low dimensional parameter estimation problem on the selected support. Our refitting estimator is defined as

$$\overline{\boldsymbol{\beta}}^r := \frac{1}{k}\sum_{j=1}^k (X_{\widehat{S}}^{(j)T}X_{\widehat{S}}^{(j)})^{-1}X_{\widehat{S}}^{(j)T}\boldsymbol{Y}^{(j)}, \tag{A.2}$$

where $\widehat{S} := \{j : |\overline{\beta}_j^d| > 2C\sqrt{\log d/n}\}$ and C is the same constant as in (4.1).

Corollary A.3. Suppose $\beta^*_{\min} > 2C\sqrt{\log d/n}$, where $\beta^*_{\min} := \min_{1\le j\le d}|\beta^*_j|$ and C is the same constant as in (4.1). Define the full sample oracle estimator as $\widehat{\boldsymbol{\beta}}^o = (X_S^T X_S)^{-1}X_S^T\boldsymbol{Y}$, where S is the true support of $\boldsymbol{\beta}^*$. If $k = O(\sqrt{n/(s^2\log d)})$, then for sufficiently large n and d we have

$$\text{(A.3)} \qquad \|\overline{\boldsymbol{\beta}}^r - \widehat{\boldsymbol{\beta}}^o\|_2 = O_{\mathbb{P}}\Big(\frac{\sqrt{k}(s\vee\log n)}{n}\Big), \quad \|\overline{\boldsymbol{\beta}}^r - \boldsymbol{\beta}^*\|_2 = O_{\mathbb{P}}\big(\sqrt{s/n}\big).$$

We see from Corollary A.3 that $\overline{\boldsymbol{\beta}}^r$ achieves the oracle rate when the minimum signal strength is not too weak and the number of subsamples k is not too large.

APPENDIX B: THE LOW-DIMENSIONAL GENERALIZED LINEAR MODEL

The next theorem quantifies the gap between $\overline{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}}$, where $\overline{\boldsymbol{\beta}}$ is the average of subsampled GLM estimators and $\widehat{\beta}$ is the full sample GLM estimator. In Theorem B.1, $\|\overline{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}\|_2$ is the distance between the divide and conquer estimator and the full sample estimator, while $\|\overline{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2$ is the estimation error on each machine.

Theorem B.1. Under Condition 3.6, if $k = O(\sqrt{n}/(d\vee\log n))$, then we have for sufficiently large d and n,

$$\text{(B.1)} \qquad \|\overline{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}\|_2 = O_{\mathbb{P}}\Big(\frac{k\sqrt{d}(d\vee\log n)}{n}\Big), \quad \|\overline{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 = O_{\mathbb{P}}\big(\sqrt{d/n}\big).$$

Remark B.2. In analogy to Theorem A.1, by constraining the growth rate of the number of subsamples according to $k = o\big(\sqrt{n}/(d\vee\log n)\big)$, the error incurred by the divide and conquer procedure, i.e., $\|\overline{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}\|_2$ decays at a faster rate than that of the statistical error of the full sample estimator $\widehat{\boldsymbol{\beta}}$.

We notice a recent independent work Liu and Ihler (2014) on distributed estimation under curved exponential families with fixed dimensions. They propose a KL-divergence-based combination method to aggregate MLEs from multiple data repositories and show that it can achieve the best possible approximation to the global MLE given the entire dataset. In the future work, it will be interesting to extend their approach to the GLM setting and characterize the statistical error rate of the correspondent distributed estimator.

The less stringent scaling of k in the low dimensional linear model relative to the generalized linear model comes from the fact that the Hessian matrix depends on the estimator of $\boldsymbol{\beta}^*$ in the GLM. This results in a larger variance relative to the linear model. Figures 3(A) and 4(A) indicate that the deduced scaling is sharp for both cases.

As in the linear model, Lemma 4.6 together with Theorem B.1 allow us to study the theoretical properties of a refitting estimator for the high dimensional GLM. Estimation on the estimated support set is again a low dimensional problem, thus the d of Theorem B.1 corresponds to the s of Lemma 4.6 in this refitting setting. The refitted GLM estimator is defined as

$$\overline{\boldsymbol{\beta}}^r = \frac{1}{k}\sum_{j=1}^{k} \widehat{\boldsymbol{\beta}}^r(\mathcal{D}_j), \tag{B.2}$$

where $\widehat{\boldsymbol{\beta}}^r(\mathcal{D}_j) = \operatorname{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^d, \boldsymbol{\beta}_{\widehat{S}^c}=0}\, \ell_{n_k}^{(j)}(\boldsymbol{\beta})$ and $\widehat{S} := \{j : |\overline{\boldsymbol{\beta}}^d_j| > 2C\sqrt{\log d/n}\}$. The following corollary quantifies the statistical rate of $\overline{\boldsymbol{\beta}}^r$.

Corollary B.3. Suppose $\beta^*_{\min} > 2C\sqrt{\log d/n}$, where $\beta^*_{\min} := \min_{1\le j\le d} |\beta^*_j|$ and C is the same constant as Lemma 4.6. Define the full sample oracle estimator as $\widehat{\boldsymbol{\beta}}^o = \operatorname{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^d, \boldsymbol{\beta}_{S^c}=0}\, \ell_n(\boldsymbol{\beta})$, where S is the true support of $\boldsymbol{\beta}^*$. If $k = O\big(\sqrt{n/((s\vee s_1)^2\log d)}\big)$, then for sufficiently large n and d we have

$$\|\overline{\boldsymbol{\beta}}^r - \widehat{\boldsymbol{\beta}}^o\|_2 = O_{\mathbb{P}}\Big(\frac{k\sqrt{s}(s\vee \log n)}{n}\Big), \quad \|\overline{\boldsymbol{\beta}}^r - \boldsymbol{\beta}^*\|_2 = O_{\mathbb{P}}\big(\sqrt{s/n}\big). \tag{B.3}$$

We thus see that $\overline{\boldsymbol{\beta}}^r$ achieves the oracle rate when the minimum signal strength is not too weak and the number of subsamples k is not too large.

APPENDIX C: SIMULATION FOR THE LOW-DIMENSIONAL LINEAR MODEL

All $n\times d$ entries of the design matrix X are generated as i.i.d. standard normal random variables and the errors $\{\varepsilon_i\}_{i=1}^n$ are i.i.d. standard normal as well. The true regression vector $\boldsymbol{\beta}^*$ satisfies $\beta^*_j = 10/\sqrt{d}$ for $j = 1,\ldots,d/2$ and $\beta^*_j = -10/\sqrt{d}$ for $j > d/2$, which guarantees that $\|\boldsymbol{\beta}^*\|_2 = 10$. Then we generate the response variable $\{Y_i\}_{i=1}^n$ according to the model (3.2). Denote the full sample ordinary least-squares estimator and the divide and conquer estimator by $\widehat{\boldsymbol{\beta}}$ and $\overline{\boldsymbol{\beta}}$ respectively. Figure 4(A) illustrates the change in the ratio $\|\overline{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}\|_2/\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2$ as the sample size increases, where k assumes three different growth rates and $d = \sqrt{n}/2$. Figure 4(B) focuses on the relationship between the statistical error of $\overline{\boldsymbol{\beta}}$ and $\log k$ under three different scalings of n and d. All the data points are obtained based on average over 100 Monte Carlo replications.

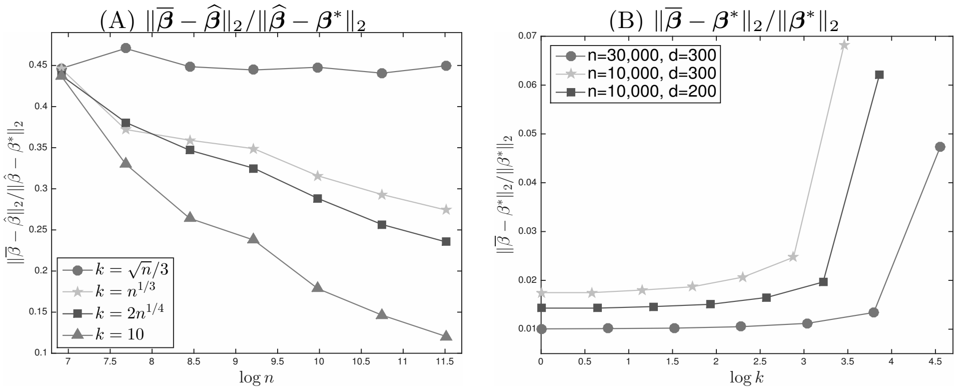


FIG 4. *(A) The ratio between the loss of the divide and conquer procedure and the statistical error of the estimator based on the whole sample with $d = \sqrt{n}/2$ and different growth rates of k. (B) Statistical error of the DC estimator against* $\log k$.

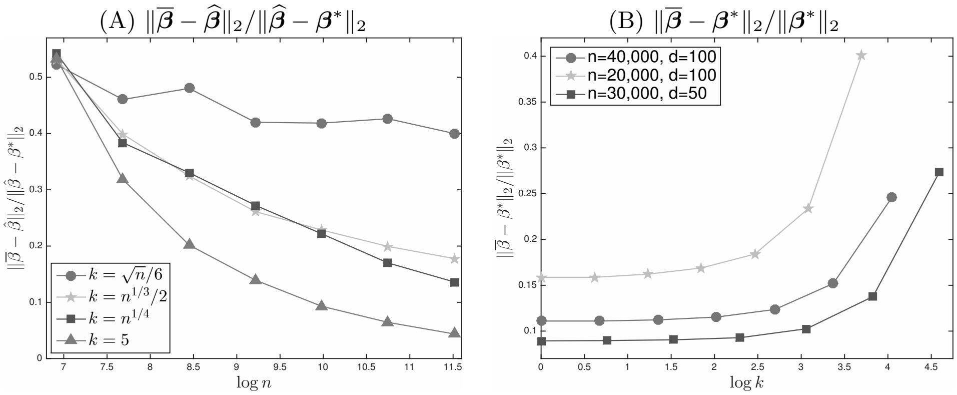
As Figure 4(A) demonstrates, when $k = O(n^{1/3})$, $O(n^{1/4})$ or $O(1)$, the ratio decreases with ever faster rates, which is consistent with the argument of Remark A.2 that the ratio goes to zero when $k = o(n/d) = o(\sqrt{n})$. When $k = O(\sqrt{n})$, however, we observe that the ratio is essentially constant, which suggests the rate we derived in Theorem A.1 is sharp. We also report the wall time of our proposed distributed approach and the naive average Lasso in Table 1. The time is computed in a similar way as the testing part. A comparison between these two approaches reveals the heavy computation incurred by debiasing. However, we observe that as the splits grow, the time consumption on individual data sample decreases since the local problem size becomes smaller, which mitigates the time complexity problem if we have a parallel computing system.

From Figure 4(B), we see that when k is not large, the statistical error of $\overline{\boldsymbol{\beta}}$ is very small because the loss incurred by the divide and conquer procedure is negligible compared to the statistical error of $\widehat{\boldsymbol{\beta}}$. However, when k is larger than a threshold, there is a surge in the statistical error, since the loss of the divide and conquer begins to dominate the statistical error of $\widehat{\boldsymbol{\beta}}$. We also notice that the larger the ratio n/d, the larger the threshold of $\log k$, which is again consistent with Remark A.2.

APPENDIX D: SIMULATION FOR THE LOW-DIMENSIONAL LOGISTIC REGRESSION

In logistic regression, given covariates $\boldsymbol{X}$, the response $Y|\boldsymbol{X} \sim \text{Ber}(\eta(\boldsymbol{X}))$, where $\text{Ber}(\eta)$ denotes the Bernoulli distribution with expectation η and

$$\eta(\boldsymbol{X}) = \frac{1}{1 + \exp(-\boldsymbol{X}^T \boldsymbol{\beta}^*)}.$$


(A) $\|\overline{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}\|_2 / \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2$

(B) $\|\overline{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 / \|\boldsymbol{\beta}^*\|_2$

FIG 5. *(A) The ratio between the loss of the divide and conquer procedure and the statistical error of the estimator based on the whole sample when $d = 20$. (B) Statistical error of the DC estimator.*

We see that $\text{Ber}(\eta(\boldsymbol{X}))$ is in exponential dispersion family canonical form (2.7) with $b(\theta) = \log(1 + e^\theta)$, $\phi = 1$ and $c(y) = 1$. The use of the canonical link,

$$\eta(\boldsymbol{X}) = \frac{1}{1 + e^{-\theta(\boldsymbol{X})}},$$

leads to the simplification $\theta(\boldsymbol{X}) = \boldsymbol{X}^T \boldsymbol{\beta}^*$.

In our Monte Carlo experiments, all $n \times d$ entries of the design matrix X are generated as i.i.d. standard normal random variables. The true regression vector $\boldsymbol{\beta}^*$ satisfies $\beta_j^* = 1/\sqrt{d}$ for $j \leq d/2$ and $\beta_j^* = -1/\sqrt{d}$ for $j > d/2$, which guarantees that $\|\boldsymbol{\beta}^*\|_2 = 1$. Finally, we generate the response variables $\{Y_i\}_{i=1}^n$ according to $\text{Ber}(\eta(\boldsymbol{X}))$. Figure 5(A) illustrates the change of the ratio $\|\overline{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}\|_2 / \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2$ as the sample size increases, where k assumes three different growths rates and $d = 20$. Figure 5(B) focuses on the relationship between the statistical error of $\overline{\boldsymbol{\beta}}$ and $\log k$ under three different scalings of n and d. All the data points are obtained based on an average over 100 Monte Carlo replications.

Figure 5 reveals similar phenomena to those revealed in Figure 4 of the previous subsection. More specifically, Figure 5(A) shows that when $k = O(n^{1/3})$, $O(n^{1/4})$ or $O(1)$, the ratio decreases with even faster rates, which is consistent with the argument of Remark B.2 that the ratio converges to zero when $k = o(\sqrt{n}/d) = o(\sqrt{n})$. When $k = O(\sqrt{n})$, however, we observe that the ratio remains essentially constant when $\log n$ is large, which suggests the rate we derived in Theorem A.1 is sharp.

As for Figure 5(B), we again observe that the statistical error of $\overline{\boldsymbol{\beta}}$ is very small when k is sufficiently small, but grows fast when k becomes large. The reasoning is the same as in the linear model, i.e. when k is large, the loss incurred by the divide and conquer procedure is non-negligible as compared with the statistical error of $\|\widehat{\boldsymbol{\beta}}\|_2$. In addition, as Figure 5(B) reveals, the larger is $\sqrt{n}/d$, the larger the threshold of k, which is again consistent with the threshold rate pointed out in Remark B.2.

APPENDIX E: AUXILIARY LEMMAS AND THEOREMS FOR TESTING

In this section, we provide the proofs of the technical lemmas and theorems for the divide and conquer hypothesis testing.

Proof of Theorem 3.3. It remains to verify the Lindeberg's Condition for (7.1). By Lemma E.1,

$$\big|\xi_{iv}^{(j)}\big| \le n^{-1/2}c_{n_k}^{-1}|\mathbf{m}_v^{(j)T}\boldsymbol{X}_i^{(j)}||\varepsilon_i^{(j)}| \le n^{-1/2}c_{n_k}^{-1}\vartheta_2|\varepsilon_i^{(j)}|,$$

where $\liminf_{n_k} c_{n_k} = c_\infty > 0$, hence the event $\big\{|\xi_{iv}^{(j)}| > \varepsilon\sigma\big\}$ is contained in the event $\big\{|\varepsilon_i^{(j)}| > \varepsilon\sigma c_{n_k}\vartheta_2^{-1}\sqrt{n}\big\}$ and we have

$$\begin{aligned}
&\frac{1}{\sigma^2}\sum_{j=1}^{k}\sum_{i\in\mathcal{I}_j}\mathbb{E}\left[(\xi_{iv}^{(j)})^2\,\mathbb{1}\big\{|\xi_{iv}^{(j)}| > \varepsilon\sigma\big\}\big|X\right]\\
\le\;& \frac{1}{\sigma^2}\sum_{j=1}^{k}\sum_{i\in\mathcal{I}_j}\mathbb{E}\left[(\xi_{iv}^{(j)})^2\,\mathbb{1}\big\{|\varepsilon_i^{(j)}| > \varepsilon\sigma c_{n_k}\vartheta_2^{-1}\sqrt{n}\big\}\big|X\right]\\
=\;& \frac{1}{\sigma^2}\frac{1}{k}\sum_{j=1}^{k}\frac{1}{n_k}\sum_{i\in\mathcal{I}_j}\frac{(\mathbf{m}_v^{(j)T}\boldsymbol{X}_i^{(j)})^2}{\mathbf{m}_v^{(j)T}\widehat{\Sigma}\mathbf{m}_v^{(j)}}\mathbb{E}\Big[(\varepsilon_i^{(j)})^2\,\mathbb{1}\big\{|\varepsilon_i^{(j)}| > \varepsilon\sigma c_{n_k}\vartheta_2^{-1}\sqrt{n}\big\}\Big]\\
=\;& \frac{1}{\sigma^2}\mathbb{E}\Big[(\varepsilon_i^{(j)})^2\,\mathbb{1}\big\{|\varepsilon_i^{(j)}| > \varepsilon\sigma c_{n_k}\vartheta_2^{-1}\sqrt{n_k}\sqrt{k}\big\}\Big].
\end{aligned}$$

Taking expectation with respect to X on both sides above yields that

$$\begin{aligned}
&\frac{1}{\sigma^2}\sum_{j=1}^{k}\sum_{i\in\mathcal{I}_j}\mathbb{E}\left[(\xi_{iv}^{(j)})^2\,\mathbb{1}\big\{|\xi_{iv}^{(j)}| > \varepsilon\sigma\big\}\right]\\
&\qquad\le \frac{1}{\sigma^2}\mathbb{E}\Big[(\varepsilon_i^{(j)})^2\,\mathbb{1}\big\{|\varepsilon_i^{(j)}| > \varepsilon\sigma c_{n_k}\vartheta_2^{-1}\sqrt{n_k}\sqrt{k}\big\}\Big].
\end{aligned}$$

Let $\delta = \varepsilon\sigma c_{n_k}\vartheta_2^{-1}\sqrt{n}$. Then, for any $\eta > 0$,

$$\mathbb{E}\big[\big(\varepsilon_i^{(j)}\big)^2\,\mathbb{1}\big\{\big|\varepsilon_i^{(j)}\big| > \delta\big\}\big] \le \mathbb{E}\Big[\big(\varepsilon_i^{(j)}\big)^2\frac{\big|\varepsilon_i^{(j)}\big|^\eta}{\delta^\eta}\,\mathbb{1}\big\{\big|\varepsilon_i^{(j)}\big| > \delta\big\}\Big] \le \delta^{-\eta}\mathbb{E}\big[\big|\varepsilon_i^{(j)}\big|^{2+\eta}\big]. \tag{E.1}$$

Since $\vartheta_2 n^{-1/2} = o(1)$ by the statement of the theorem, the choice $\eta = 2$ delivers

$$\frac{1}{\sigma^2} \lim_{k\to\infty} \lim_{n_k\to\infty} \sum_{j=1}^{k} \sum_{i\in\mathcal{I}_j} \mathbb{E}\left[(\xi_{iv}^{(j)})^2 \, \mathbb{1}\{ |\xi_{iv}^{(j)}| > \varepsilon\sigma \} \right]$$

$$\le \lim_{k\to\infty} \lim_{n_k\to\infty} k^{-1} n_k^{-1} \vartheta_2 c_{n_k}^{-2} \varepsilon^{-2} \sigma^{-2} \mathbb{E}\left((\varepsilon_i^{(j)})^4 \right) = 0 \tag{E.2}$$

by the bounded forth moment assumption. By the law of iterated expectations, all conditional results hold in unconditional form as well. Hence, $\overline{V}_n \rightsquigarrow N(0,\sigma^2)$ by the Lindeberg-Feller central limit theorem. □

Proof of Corollary 3.9. We verify (A5)-(A9) of Lemma E.8 in the Supplementary Material. (A5) is satisfied because $\widetilde{\Theta}_{vv}$ is consistent under the required scaling by the statement of the corollary. (A6) is satisfied by Condition 3.7. To verify (A7), first note that $\nabla \ell_i(\boldsymbol{\beta}^*) = (b'(\boldsymbol{X}_i^T \boldsymbol{\beta}^*) - Y_i)\boldsymbol{X}_i$. According to Lemma E.2 in the Supplementary Material, we know that conditional on X, $b'(\boldsymbol{X}_i^T\boldsymbol{\beta}^*) - Y_i$ is a sub-Gaussian random variable. Therefore Lemma F.6 in the Supplementary Material delivers

$$\mathbb{P}\left(\| \frac{1}{n} \sum_{j=1}^{k} \sum_{i\in\mathcal{I}_j} \nabla \ell_i(\boldsymbol{\beta}^*) \|_\infty > t \,|\, X \right) \le d \exp\left(1 - \frac{ct^2}{nM^2} \right),$$

which implies that with probability $1 - c/d$,

$$\| \sum_{j=1}^{k} \sum_{i\in\mathcal{I}_j} \nabla \ell_i(\boldsymbol{\beta}^*) \|_\infty = C\sqrt{n \log d} \tag{E.3}$$

It only remains to verify (A8). Let $\xi_{iv}^{(j)} = \boldsymbol{\Theta}_v^{*T} \nabla \ell_i^{(j)}(\boldsymbol{\beta}^*)/\sqrt{n\Theta_{vv}^*}$. By the definition of the log likelihood,

$$\mathbb{E}[\xi_{iv}^{(j)}] = \frac{\boldsymbol{\Theta}_v^{*T} \mathbb{E}[\nabla \ell_i^{(j)}(\boldsymbol{\beta}^*)]}{(n\Theta_{vv}^*)^{1/2}} = 0$$

and by independence of $\{(Y_i, \boldsymbol{X}_i)\}_{i=1}^n$,

$$\mathrm{Var}\big(\sum_{j=1}^{k}\sum_{i\in\mathcal{I}_j} \xi_{iv}^{(j)}\big) = \sum_{j=1}^{k}\sum_{i\in\mathcal{I}_j} \mathrm{Var}\big(\xi_{iv}^{(j)}\big) = \sum_{j=1}^{k}\sum_{i\in\mathcal{I}_j} \mathbb{E}[(\xi_{iv}^{(j)})^2]$$

$$= \frac{1}{n}\sum_{i=1}^{n} (\Theta_{vv}^*)^{-1} \boldsymbol{\Theta}_v^{*T} \mathbb{E}\big[\big(\nabla \ell_i(\boldsymbol{\beta}^*)\big)\big(\nabla \ell_i(\boldsymbol{\beta}^*)\big)^T \big] \boldsymbol{\Theta}_v^*$$

$$= \frac{1}{n}\sum_{i=1}^{n} (\Theta_{vv}^*)^{-1} [\Theta^* J^* \Theta^*]_{vv} = 1.$$

By Condition 3.6, $\theta_{\min} > 0$, the event $\{|\xi_{iv}^{(j)}| > \varepsilon\}$ coincides with the event $\big\{\big|\boldsymbol{\Theta}_v^{*T}\nabla\ell_i(\boldsymbol{\beta}^*)\big| > \varepsilon\sqrt{\theta_{\min} n}\big\} = \big\{\big|\boldsymbol{\Theta}_v^{*T}\boldsymbol{X}_i(Y_i - b'(\boldsymbol{X}_i^T\boldsymbol{\beta}^*))\big| > \varepsilon\sqrt{\theta_{\min} n}\big\}$. Furthermore, since $\big|\boldsymbol{\Theta}_v^{*T}\boldsymbol{X}_i\big| \leq M$ by Condition 3.7, this event is contained in the event $\big\{\big|Y_i - b'(\boldsymbol{X}_i^T\boldsymbol{\beta}^*)\big| > \delta\big\}$, where $\delta = \varepsilon\sqrt{\theta_{\min} n}/M$. By an analogous calculation to that of equation (E.1) in the Supplementary Material, we have

$$\mathbb{E}\Big[\big(Y_i - b'(\boldsymbol{X}_i^T\boldsymbol{\beta}^*)\big)^2 \mathbb{1}\{|Y_i - b'(\boldsymbol{X}_i^T\boldsymbol{\beta}^*)| > \delta\}|X\Big] \leq \delta^{-\eta}\mathbb{E}\big[\big(Y_i - b'(\boldsymbol{X}_i^T\boldsymbol{\beta}^*)\big)^{2+\eta}|X\big].$$

Hence, setting $\eta = 2$ and noting that $\mathbb{E}\big[(Y_i - b'(\boldsymbol{X}_i^T\boldsymbol{\beta}^*))^{2+\eta}|X\big] \leq C\sqrt{2+\eta}\phi U_2$ by Lemma E.2 in the Supplementary Material, it follows that

$$\begin{aligned}
&\lim_{k\to\infty}\lim_{n_k\to\infty}\sum_{j=1}^{k}\sum_{i\in\mathcal{I}_j}\mathbb{E}\big[(\xi_{i,v}^{(j)})^2\,\mathbb{1}\{|\xi_{i,v}^{(j)}| > \varepsilon\}\big]\\
\leq\ & (\theta_{\min})^{-1}\lim_{k\to\infty}\lim_{n_k\to\infty} n^{-1}\sum_{j=1}^{k}\sum_{i\in\mathcal{I}_j}\boldsymbol{\Theta}_v^{*T}\mathbb{E}[\boldsymbol{X}_i\boldsymbol{X}_i^T]\boldsymbol{\Theta}_v^*\delta^{-2}\\
\leq\ & (\theta_{\min})^{-1}\lim_{k\to\infty}\lim_{n_k\to\infty} M^3 s_1^2/(n\varepsilon^2\theta_{\min}) = 0, \qquad \text{(E.4)}
\end{aligned}$$

where the last inequality follows because $\|\Sigma\|_{\max} = \|\mathbb{E}[\boldsymbol{X}_i\boldsymbol{X}_i^T]\|_{\max} < M^2$ by Condition 3.6. Similarly, we have for any $\varepsilon > 0$,

$$\varepsilon^{-3}\lim_{k\to\infty}\lim_{n_k\to\infty}\sum_{j=1}^{k}\sum_{i\in\mathcal{I}_j}\mathbb{E}\big[(\xi_{i,v}^{(j)})^3\,\mathbb{1}\{|\xi_{i,v}^{(j)}| > \varepsilon\}\big] = 0.$$

Applying the self-normalized Berry-Essen inequality, we complete the proof of this corollary. □

Lemma E.1. Under Condition 3.2, $\big(\mathbf{m}_v^{(j)T}\widehat{\Sigma}\mathbf{m}_v^{(j)}\big)^{-1/2} \geq c_{n_k}$ for any $j \in \{1,\ldots,k\}$ and for any $v \in \{1,\ldots,d\}$, where c_{n_k} satisfies $\liminf_{n_k\to\infty} c_{n_k} = c_\infty > 0$.

PROOF. The proof appears in the proof of Lemma B1 of Zhao et al. (2014b). □

Lemma E.2. Under the GLM (2.7), we have

$$\mathbb{E}\exp(t(Y - \mu(\theta))) = \exp(\phi^{-1}(b(\theta + t\phi) - b(\theta) - \phi t b'(\theta))),$$

and typically when there exists $U > 0$ such that $b''(\theta) < U$ for all $\theta \in \mathbb{R}$, we will have

$$\mathbb{E}\exp(t(Y - \mu(\theta))) \leq \exp\left(\frac{\phi U t^2}{2}\right),$$

which implies that Y is a sub-Gaussian random variable with variance proxy ϕU.

Proof.

$$\begin{aligned}
&\mathbb{E}\exp\left(t(Y-\mu(\theta))\right) \\
&= \int_{-\infty}^{+\infty} c(y)\exp\left(\frac{y\theta-b(\theta)}{\phi}\right)\exp(t(y-\mu(\theta)))dy \\
&= \int_{-\infty}^{+\infty} c(y)\exp\left(\frac{(\theta+t\phi)y-(b(\theta)+\phi t b'(\theta))}{\phi}\right)dy \\
&= \int_{-\infty}^{+\infty} c(y)\exp\left(\frac{(\theta+t\phi)y-b(\theta+t\phi)+b(\theta+t\phi)-(b(\theta)+\phi t b'(\theta))}{\phi}\right)dy \\
&= \exp\left(\phi^{-1}(b(\theta+t\phi)-b(\theta)-\phi t b'(\theta))\right).
\end{aligned}$$

When $b''(\theta) < U$. the mean value theorem gives

$$\mathbb{E}\exp\left(t(Y-\mu(\theta))\right) = \exp\left(\frac{b''(\widetilde{\theta})\phi^2 t^2}{2\phi}\right) \le \exp\left(\frac{\phi U t^2}{2}\right).$$

□

Proof of Lemma 3.4. We first show that, for any $j \in \{1,\dots,k\}$, $|\widehat{\sigma}^2(\mathcal{D}_j) - \sigma^2| = o_{\mathbb{P}}(k^{-1})$. To this end, letting

$$\widehat{\varepsilon}_i = Y_i^{(j)} - \boldsymbol{X}_i^{(j)T}\widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j) = Y_i^{(j)} - \boldsymbol{X}_i^{(j)T}\boldsymbol{\beta}^* - \boldsymbol{X}_i^{(j)T}\big(\widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j) - \boldsymbol{\beta}^*\big),$$

we write

$$|\widehat{\sigma}^2(\mathcal{D}_j) - \sigma^2| = \Big|\frac{1}{n_k}\sum_{i\in\mathcal{I}_j}\widehat{\varepsilon}_i^2 - \sigma^2\Big| \le \Delta_1^{(j)} + 2\Delta_2^{(j)} + \Delta_3^{(j)},$$

$$\begin{aligned}
\Delta_1^{(j)} &:= \big|\frac{1}{n_k}\sum_{i\in\mathcal{I}_j}\varepsilon_i^2 - \sigma^2\big|,\ \Delta_2^{(j)} := \big|\big(\widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j) - \boldsymbol{\beta}^*\big)\big(\frac{1}{n_k}\sum_{i\in\mathcal{I}_j}\boldsymbol{X}_i^{(j)}\varepsilon_i^{(j)}\big)\big| \text{ and} \\
\Delta_3^{(j)} &:= \big|\big(\widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j) - \boldsymbol{\beta}^*\big)^T\Big(\frac{1}{n_k}\sum_{i\in\mathcal{I}_j}\boldsymbol{X}_i^{(j)}\boldsymbol{X}_i^{(j)T}\Big)\big(\widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j) - \boldsymbol{\beta}^*\big)\big| \\
&= \big\|X^{(j)}\big(\widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j) - \boldsymbol{\beta}^*\big)\big\|_2^2/n_k = O_{\mathbb{P}}(\lambda^2 s)
\end{aligned}$$

by Theorem 6.1 of Bühlmann and van de Geer (2011). Hence, with $\lambda = C\sigma^2\sqrt{k\log d/n}$, $\Delta_3^{(j)} = o_{\mathbb{P}}(1)$ for $k = o\big((s\log d)^{-1}n\big)$, a fortiori for $k =$

$o\big((s\log d)^{-1}\sqrt{n}\big)$. Letting

$$\begin{aligned}\Delta_{21}^{(j)} &= \big\|\widehat{\boldsymbol{\beta}}^{\lambda}(\mathcal{D}_j)-\boldsymbol{\beta}^*\big\|_1\Big\|\frac{1}{n_k}\sum_{i\in\mathcal{I}_j}\boldsymbol{X}_i^{(j)}\varepsilon_i^{(j)}-\mathbb{E}[\boldsymbol{X}_i^{(j)}\varepsilon_i^{(j)}]\Big\|_\infty,\\ \Delta_{22}^{(j)} &= \big\|\widehat{\boldsymbol{\beta}}^{\lambda}(\mathcal{D}_j)-\boldsymbol{\beta}^*\big\|_1\big\|\mathbb{E}[\boldsymbol{X}_i^{(j)}\varepsilon_i^{(j)}]\big\|_\infty.\end{aligned}$$

We obtain the bound

$$\begin{aligned}\Delta_2^{(j)} &= \Big|\big(\widehat{\boldsymbol{\beta}}^{\lambda}(\mathcal{D}_j)-\boldsymbol{\beta}^*\big)\Big(\big(\frac{1}{n_k}\sum_{i\in\mathcal{I}_j}\boldsymbol{X}_i^{(j)}\varepsilon_i^{(j)}-\mathbb{E}[\boldsymbol{X}_i^{(j)}\varepsilon_i^{(j)}]\big)+\mathbb{E}[\boldsymbol{X}_i^{(j)}\varepsilon_i^{(j)}]\Big)\Big|\\ &\le \Delta_{21}^{(j)}+\Delta_{22}^{(j)}.\end{aligned}$$

By the statement of the Lemma, $\mathbb{E}\big[\boldsymbol{X}_i^{(j)}\varepsilon_i^{(j)}\big]=\mathbb{E}\big[\boldsymbol{X}_i^{(j)}\mathbb{E}[\varepsilon_i^{(j)}|\boldsymbol{X}_i^{(j)}]\big]=0$, hence $\Delta_{22}^{(j)}=0$, while by the central limit theorem and Theorem 6.1 of Bühlmann and van de Geer (2011),

$$\Delta_{21}^{(j)}\le O_{\mathbb{P}}(\lambda s)O_{\mathbb{P}}(n_k^{-1/2}).$$

We conclude $\Delta_2^{(j)}=O_{\mathbb{P}}\big(\lambda s n_k^{-1/2}\big)$, and with $\lambda\asymp\sigma^2\sqrt{k\log d/n}$, $\Delta_2^{(j)}=o(1)$ with $k=o\big(n(s\log d)^{-2/3}\big)$, a fortiori for $k=o\big(\sqrt{n}(s\log d)^{-1}\big)$. Finally, noting that $\sigma^2=\mathbb{E}[\varepsilon_i^{(j)}]$, $\Delta_1^{(j)}=O_{\mathbb{P}}(n_k^{-1/2})=o_{\mathbb{P}}\big(1\big)$ by the central limit theorem. Combining the bounds, we obtain $|\widehat{\sigma}^2(\mathcal{D}_j)-\sigma^2|=o_{\mathbb{P}}(1)$ for any $j\in\{1,\dots,k\}$ and therefore $|\overline{\sigma}^2-\sigma^2|\le k^{-1}\sum_{j=1}^k|\widehat{\sigma}^2(\mathcal{D}_j)-\sigma^2|=o_{\mathbb{P}}(1)$. □

Lemma E.3. Under Condition 3.6, we have for any $\boldsymbol{\beta},\boldsymbol{\beta}'\in\mathbb{R}^d$ and any $i=1,\dots,n$, $\big|\ell_i''(\boldsymbol{X}_i^T\boldsymbol{\beta})-\ell_i''(\boldsymbol{X}_i^T\boldsymbol{\beta}')\big|\le K_i|\boldsymbol{X}_i^T(\boldsymbol{\beta}-\boldsymbol{\beta}')|$, where $0<K_i<\infty$.

Proof. By the canonical form of the generalized linear model (equation (2.8)),

$$\big|\ell_i''(\boldsymbol{X}_i^T\boldsymbol{\beta})-\ell_i''(\boldsymbol{X}_i^T\boldsymbol{\beta}')\big|=\big|b''(\boldsymbol{X}_i^T\boldsymbol{\beta})-b''(\boldsymbol{X}_i^T\boldsymbol{\beta}')\big|\le|b'''(\widetilde{\eta})||\boldsymbol{X}_i^T(\boldsymbol{\beta}-\boldsymbol{\beta}')|$$

by the mean value theorem, where $\widetilde{\eta}$ lies in a line segment between $\boldsymbol{X}_i^T\boldsymbol{\beta}$ and $\boldsymbol{X}_i^T\boldsymbol{\beta}'$. $|b'''(\eta)|<U_3<\infty$ by Condition 3.6 for any η, hence the conclusion follows with $K_i=U_3$ for all i. □

Lemma E.4. Under Conditions 2.6 and 2.1 (i), we have for any $\delta\in(0,1)$ such that $\delta^{-1}\ll d$,

$$\mathbb{P}\Big(\frac{1}{n}\big\|\boldsymbol{X}(\widehat{\boldsymbol{\beta}}^{\lambda}-\boldsymbol{\beta}^*)\big\|_2^2\gtrsim s\frac{\log(d/\delta)}{n}\Big)<\delta$$

PROOF. Decompose the object of interest as

$$\begin{aligned}\frac{1}{n}\|\boldsymbol{X}(\widehat{\boldsymbol{\beta}}^\lambda-\boldsymbol{\beta}^*)\|_2^2 &= (\widehat{\boldsymbol{\beta}}^\lambda-\boldsymbol{\beta}^*)^T(\widehat{\Sigma}-\Sigma)(\widehat{\boldsymbol{\beta}}^\lambda-\boldsymbol{\beta}^*)+(\widehat{\boldsymbol{\beta}}^\lambda-\boldsymbol{\beta}^*)^T\Sigma(\widehat{\boldsymbol{\beta}}^\lambda-\boldsymbol{\beta}^*)\\ &\le \|\widehat{\Sigma}-\Sigma\|_{\max}\|\widehat{\boldsymbol{\beta}}^\lambda-\boldsymbol{\beta}^*\|_1^2+\lambda_{\max}(\Sigma)\|\widehat{\boldsymbol{\beta}}^\lambda-\boldsymbol{\beta}^*\|_2^2.\end{aligned}$$

This gives rise to the tail probability bound

$$\begin{aligned}&\mathbb{P}\Big(\frac{1}{n}\|\boldsymbol{X}(\widehat{\boldsymbol{\beta}}^\lambda-\boldsymbol{\beta}^*)\|_2^2>t\Big)\\ &\le \mathbb{P}\Big(\|\widehat{\Sigma}-\Sigma\|_{\max}\|\widehat{\boldsymbol{\beta}}^\lambda-\boldsymbol{\beta}^*\|_1^2>\frac{t}{2}\Big)+\mathbb{P}\Big(\lambda_{\max}(\Sigma)\|\widehat{\boldsymbol{\beta}}^\lambda-\boldsymbol{\beta}^*\|_2^2>\frac{t}{2}\Big).\end{aligned} \tag{E.5}$$

Let $\mathcal{M}:=\big\{\|\widehat{\Sigma}-\Sigma\|_\infty\le M\big\}$. Since $\{\boldsymbol{X}_i\}_{i=1}^n$ is bounded, it is sub-Gaussian as well. Suppose $\|\boldsymbol{X}_i\|_{\psi_2}<\kappa$, then by Lemma F.3 we have,

$$\mathbb{P}(\mathcal{M}^c)\le\sum_{p,q=1}^d\mathbb{P}(|\widehat{\Sigma}_{pq}^{(j)}-\Sigma_{pq}|>M)\le d^2\exp\left(-Cn\cdot\min\Big\{\frac{M^2}{\kappa^4},\frac{M}{\kappa^2}\Big\}\right),$$

where C is a constant. Hence taking $M=n^{-1}\log(d/\delta)$,

$$\mathbb{P}(\mathcal{M}^c)\le d^2\exp\left\{-Cn\min\Big\{\frac{(\log(d/\delta))^2}{\kappa^4n^2},\frac{(\log(d/\delta))^2}{\kappa^2n}\Big\}\right\}$$

and the right hand side is less than δ for $\delta^{-1}\ll d$. Thus by Condition 2.1, the first term on the right hand side of equation (E.5) is

$$\mathbb{P}\Big(\big\|\widehat{\Sigma}-\Sigma\big\|_{\max}\big\|\widehat{\boldsymbol{\beta}}^\lambda-\boldsymbol{\beta}^*\big\|_1^2\gtrsim\frac{s\log(d/\delta)}{n}\Big)<2\delta.$$

Furthermore, by Condition 3.6 (i), the second term on the right hand side of equation (E.5) is

$$\mathbb{P}\Big(\lambda_{\max}(\Sigma)\big\|\widehat{\boldsymbol{\beta}}^\lambda-\boldsymbol{\beta}^*\big\|_2^2\gtrsim C_{\max}\frac{s\log(d/\delta)}{n}\Big)<\delta.$$

Taking t as the dominant term, $t\asymp C_{\max}n^{-1}s\log(d/\delta)$, yields the result. □

Lemma E.5. Under Condition 3.6, we have for any $i=1,\ldots,n$,

$$|b''(\boldsymbol{X}_i^T\boldsymbol{\beta}_1)-b''(\boldsymbol{X}_i^T\boldsymbol{\beta}_2)|\le MU_3\|\boldsymbol{\beta}_1-\boldsymbol{\beta}_2\|_1,$$

and if we consider the sub-Gaussian design instead, we have

$$\mathbb{P}\left(|b''(\boldsymbol{X}_i^T\boldsymbol{\beta}_1)-b''(\boldsymbol{X}_i^T\boldsymbol{\beta}_2)|\ge hU_3\|\boldsymbol{\beta}_1-\boldsymbol{\beta}_2\|_1\right)\le nd\exp\left(1-\frac{Ch^2}{s_1^2}\right).$$

PROOF. For the bounded design, by Condition 3.6 (iii), we have

$$\begin{aligned}|b''(\boldsymbol{X}_i^T\boldsymbol{\beta}_1) - b''(\boldsymbol{X}_i^T\boldsymbol{\beta}_2)| &\le U_3|\boldsymbol{X}_i^T(\boldsymbol{\beta}_1-\boldsymbol{\beta}_2)| \le U_3\|\boldsymbol{X}_i\|_{\max}\|\boldsymbol{\beta}_1-\boldsymbol{\beta}_2\|_1\\ &\le MU_3\|\boldsymbol{\beta}_1-\boldsymbol{\beta}_2\|_1.\end{aligned}$$

For the sub-Gaussian design, denote the event $\{\max_{1\le i\le n, 1\le j\le d}|X_{ij}|\le h\}$ by $\mathcal{C}$, where κ is a positive constant. Then it follows that,

$$\mathbb{P}\left(\mathcal{C}^c\right) \le nd\exp\left(1-\frac{Ch^2}{s_1^2}\right),$$

where C is a constant. Since on the event $\mathcal{C}$, $|b''(\boldsymbol{X}_i^T\boldsymbol{\beta}_1) - b''(\boldsymbol{X}_i^T\boldsymbol{\beta}_2)| \le hU_3\|\boldsymbol{\beta}_1-\boldsymbol{\beta}_2\|_1$, we reach the conclusion. □

Remark E.6. For the sub-Gaussian design, in order to let the tail probability go to zero, $h \gg \log((n\vee d))$.

Lemma E.7. *Suppose, for any $k \ll d$ satisfying $k = o\big(((s\vee s_1)logd)^{-1}\sqrt{n}\big)$, the following conditions are satisfied. (A1) $\mathbb{P}\left(n_k^{-1}\|X^{(j)}\widehat{\boldsymbol{\Theta}}^{(j)}\|_{\max} \ge H\right) \le \xi$, where H is a constant and $\xi = o(k^{-1})$. (A2) For any $\boldsymbol{\beta},\boldsymbol{\beta}' \in \mathbb{R}^d$ and for any $i \in \{1,\ldots,n\}$, $\big|\ell_i''(\boldsymbol{X}_i^T\boldsymbol{\beta}) - \ell_i''(\boldsymbol{X}_i^T\boldsymbol{\beta}')\big| \le K_i\big|\boldsymbol{X}_i^T(\boldsymbol{\beta}-\boldsymbol{\beta}')\big|$ with $\mathbb{P}(K_i > h) \le \psi$ for $\psi = o(k^{-1})$ and $h = O(1)$. (A3) $\mathbb{P}\Big(n_k^{-1}\|X^{(j)}(\widehat{\boldsymbol{\beta}}^\lambda - \boldsymbol{\beta}^*)\|_2^2 \gtrsim n^{-1}sk\log(d/\delta)\Big) < \delta$. (A4) $\mathbb{P}\Big(\max_{1\le v\le d}\Big|\big(\widehat{\boldsymbol{\Theta}}_v^{(j)T}\nabla^2\ell_{n_k}^{(j)}(\widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j)) - \boldsymbol{e}_v\big)\big(\widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j)-\boldsymbol{\beta}^*\big)\Big| \gtrsim n^{-1}sk\log(d/\delta)\Big) < \delta$. Then*

$$\overline{\beta}_v^d - \beta_v^* = -\frac{1}{k}\sum_{j=1}^k \widehat{\boldsymbol{\Theta}}_v^{(j)T}\nabla\ell_{n_k}^{(j)}(\boldsymbol{\beta}^*) + o_{\mathbb{P}}(n^{-1/2}).$$

for any $1 \le v \le d$.

PROOF OF LEMMA E.7. $\overline{\beta}_v^d - \beta_v^* = k^{-1}\sum_{j=1}^k\big(\widehat{\beta}_v(\mathcal{D}_j) - \beta_v^*\big)$. By the definition of $\widehat{\boldsymbol{\beta}}^d(\mathcal{D}_j)$,

$$\widehat{\beta}_v^d(\mathcal{D}_j) - \beta_v^* = \widehat{\beta}_v^\lambda(\mathcal{D}_j) - \beta_v^* - \widehat{\boldsymbol{\Theta}}_v^{(j)T}\nabla\ell_{n_k}^{(j)}(\widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j)).$$

Consider a mean value expansion of $\nabla\ell_{n_k}^{(j)}\big(\widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j)\big)$ around $\boldsymbol{\beta}^*$:

$$\nabla\ell_{n_k}^{(j)}\big(\widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j)\big) = \nabla\ell_{n_k}^{(j)}(\boldsymbol{\beta}^*) + \nabla^2\ell_{n_k}^{(j)}(\boldsymbol{\beta}_\alpha)\big(\widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j)-\boldsymbol{\beta}^*\big),$$

where $\boldsymbol{\beta}_\alpha = \alpha\widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j) + (1-\alpha)\boldsymbol{\beta}^*$, $\alpha \in [0,1]$. So

$$\frac{1}{k}\sum_{j=1}^{k} \widehat{\beta}_v^d(\mathcal{D}_j) - \beta_v^* = -\frac{1}{k}\sum_{j=1}^{k} \widehat{\boldsymbol{\Theta}}_v^{(j)T}\nabla\ell_{n_k}^{(j)}(\boldsymbol{\beta}^*) - \Delta,$$

where $\Delta = \frac{1}{k}\sum_{j=1}^{k}\big(\widehat{\boldsymbol{\Theta}}_v^{(j)T}\nabla^2\ell_{n_k}^{(j)}(\boldsymbol{\beta}_\alpha) - \boldsymbol{e}_v\big)\big(\widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j) - \boldsymbol{\beta}^*\big)$. Note that $|\Delta| \le \frac{1}{k}\sum_{j=1}^{k}\big(|\Delta_1^{(j)}| + |\Delta_2^{(j)}|\big)$, where

$$\big|\Delta_1^{(j)}\big| = \Big|\big(\widehat{\boldsymbol{\Theta}}_v^{(j)T}\nabla^2\ell_{n_k}^{(j)}\big(\widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j)\big) - \boldsymbol{e}_v\big)\big(\widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j) - \boldsymbol{\beta}^*\big)\Big|.$$

By (A4) of the lemma, for $t \asymp n^{-1}sk\log(d/\delta)$,

$$\mathbb{P}\Big(|\sum_{j=1}^{k}\Delta_1^{(j)}| > kt\Big) \le \mathbb{P}\Big(\cup_{j=1}^{k}|\Delta_1^{(j)}| > t\Big) \le \sum_{j=1}^{k}\mathbb{P}(|\Delta_1^{(j)}| > t) < k\delta.$$

Substituting $\delta = o(k^{-1})$ in the expression for t and noting that $k \ll d$, we obtain $k^{-1}\sum_{j=1}^{k}\Delta_1^{(j)} = o_{\mathbb{P}}(n^{-1/2})$ for $k = o\big((s\log d)^{-1}\sqrt{n}\big)$. By (A2),

$$\begin{aligned}
\big|\Delta_2^{(j)}\big| &= \Big|\widehat{\boldsymbol{\Theta}}_v^{(j)T}\big(\nabla^2\ell_{n_k}^{(j)}(\boldsymbol{\beta}_\alpha) - \nabla^2\ell_{n_k}^{(j)}(\widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j))\big)\big(\widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j) - \boldsymbol{\beta}^*\big)\Big| \\
&= \Big|\frac{1}{n_k}\sum_{i\in\mathcal{I}_j}\widehat{\boldsymbol{\Theta}}_v^{(j)T}\boldsymbol{X}_i\boldsymbol{X}_i^T\big(\widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j) - \boldsymbol{\beta}^*\big)\big(\ell_i''(\boldsymbol{X}_i^T\boldsymbol{\beta}_\alpha) - \ell_i''(\boldsymbol{X}_i^T\widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j))\big)\Big| \\
&\le \Big(\max_{1\le i\le n} K_i\Big)\Big(\frac{1}{n_k}\|X^{(j)}\widehat{\Theta}^{(j)}\|_{\max}\Big)\Big\|\frac{1}{n_k}X^{(j)}\big(\widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j) - \boldsymbol{\beta}^*\big)\Big\|_2^2,
\end{aligned}$$

therefore by (A1) and (A3) of the lemma, for $t \asymp n^{-1}sk\log(d/\delta)$,

$$\mathbb{P}\Big(|\sum_{j=1}^{k}\Delta_2^{(j)}| > kt\Big) \le \mathbb{P}\Big(\cup_{j=1}^{k}|\Delta_2^{(j)}| > t\Big) \le \sum_{j=1}^{k}\mathbb{P}(|\Delta_2^{(j)}| > t) < k(\psi + \delta + \xi).$$

Substituting $\delta = o(k^{-1})$ in the expression for t and noting that $k \ll d$, we obtain $k^{-1}\sum_{j=1}^{k}\Delta_2^{(j)} = o_{\mathbb{P}}(n^{-1/2})$ for $sk\log(d/\delta) = o(\sqrt{n})$, i.e. for $k = o\big((s\log d)^{-1}\sqrt{n}\big)$. Combining these two results delivers $\Delta = o_{\mathbb{P}}(n^{-1/2})$ for $k = o\big((s\log d)^{-1}\sqrt{n}\big)$. □

Lemma E.8. Suppose, in addition to Conditions (A1)-(A5) of Lemma E.7, (A5) $\big|\widetilde{\Theta}_{vv} - \Theta_{vv}^*\big| = o_{\mathbb{P}}(1)$ for all $v \in \{1,\ldots,d\}$; (A6) $1/\Theta_{vv}^* = O(1)$ for all $v \in \{1,\ldots,d\}$; (A7) $\|\sum_{1\le j\le k}\sum_{i\in\mathcal{I}_j}\nabla\ell_i(\boldsymbol{\beta}^*)\|_\infty = O_{\mathbb{P}}(\sqrt{n\log d})$; (A8)

For each $v \in \{1, \dots, d\}$, letting $\xi_{iv}^{(j)} = \mathbf{\Theta}_v^{*T} \nabla \ell_i^{(j)}(\boldsymbol{\beta}^*)/\sqrt{n\Theta_{vv}^*}$, $\mathbb{E}[\xi_{iv}^{(j)}] = 0$, $\mathrm{Var}\big(\sum_{j=1}^k \sum_{i\in\mathcal{I}_j} \xi_{iv}^{(j)}\big) = 1$ and, for all $\varepsilon > 0$,

$$\lim_{k\to\infty} \lim_{n_k\to\infty} \sum_{j=1}^k \sum_{i\in\mathcal{D}_j} \mathbb{E}\big[(\xi_{iv}^{(j)})^2\, \mathbb{1}\{|\xi_{iv}^{(j)}| > \varepsilon\}\big] = 0. \tag{E.6}$$

Then under $H_0 : \beta_v^* = \beta_v^H$, taking $k = o(((s \vee s_1)\log d)^{-1}\sqrt{n})$ delivers $\overline{S}_n \rightsquigarrow N(0,1)$, where $\overline{S}_n$ is defined in equation (3.14).

Proof. Rewrite equation (3.14) as

$$\begin{aligned}
\overline{S}_n &= \sqrt{n}\frac{1}{k}\sum_{j=1}^k \left[\frac{\widehat{\beta}_v^d - \beta_v^H}{(\Theta_{vv}^*)^{1/2}} + \frac{\widehat{\beta}_v^d - \beta_v^H}{(\Theta_{vv}^*)^{1/2}}\left(\frac{(\Theta_{vv}^*)^{1/2}}{\big[\widehat{\Theta}^{(j)}\widehat{H}^{(j)}\widehat{\Theta}^{(j)T}\big]_{vv}^{1/2}} - 1\right)\right] \\
&= \sum_{j=1}^k \sum_{i\in\mathcal{I}_j} \big(\Delta_{1,i}^{(j)} + \Delta_{2,i}^{(j)}\big), \qquad \text{where} \qquad \text{(E.7)}
\end{aligned}$$

$$\Delta_{1,i}^{(j)} = \frac{\widehat{\mathbf{\Theta}}_v^{(j)T} \nabla \ell_i^{(j)}(\boldsymbol{\beta}^*)}{(n\Theta_{vv}^*)^{1/2}}, \qquad \Delta_{2,i}^{(j)} = \frac{\widehat{\mathbf{\Theta}}_v^{(j)T} \nabla \ell_i^{(j)}(\boldsymbol{\beta}^*)}{(n\Theta_{vv}^*)^{1/2}}\left(\frac{(\Theta_{vv}^*)^{1/2}}{\overline{\Theta}_{vv}^{1/2}} - 1\right).$$

Further decomposing the first term, we have

$$\sum_{j=1}^k \sum_{i\in\mathcal{I}_j} \Delta_{1,i}^{(j)} = \sum_{j=1}^k \sum_{i\in\mathcal{I}_j} \xi_{i,v}^{(j)} + \Delta, \quad \text{where} \quad \Delta = \sum_{j=1}^k \sum_{i\in\mathcal{I}_j} \big(\widehat{\mathbf{\Theta}}_v^{(j)} - \mathbf{\Theta}_v^*\big)^T \frac{\nabla \ell_i(\boldsymbol{\beta}^*)}{(n\Theta_{vv}^*)^{1/2}}$$

and $\sum_{j=1}^k \sum_{i\in\mathcal{I}_j} \xi_{i,v}^{(j)} \rightsquigarrow N(0,1)$ by the Lindeberg-Feller central limit theorem. Then by Hölder's inequality, Condition 3.7 and Assumption (A6) and (A7),

$$\begin{aligned}
|\Delta| &\le \max_{1\le j\le k} \big\|\widehat{\mathbf{\Theta}}_v^{(j)} - \mathbf{\Theta}_v^*\big\|_1 \frac{\big\|\sum_{j=1}^k \sum_{i\in\mathcal{I}_j} \nabla \ell_i(\boldsymbol{\beta}^*)\big\|_\infty}{(n\Theta_{vv}^*)^{1/2}} \\
&= O_{\mathbb{P}}\Big(s_1\sqrt{\frac{k\log d}{n}}\Big) O_{\mathbb{P}}(\sqrt{\log d}) = o_{\mathbb{P}}(1),
\end{aligned}$$

where the last equation holds with the choice of $k = o((s_1 \log d)^{-1}\sqrt{n})$. Letting $\overline{\Delta}^{(j)} = (\Theta_{vv}^*)^{1/2} - \overline{\Theta}_{vv}^{1/2}$ we have

$$\begin{aligned}
\sum_{j=1}^k \sum_{i\in\mathcal{I}_j} \Delta_{2,i}^{(j)} &= \sum_{j=1}^k \sum_{i\in\mathcal{I}_j} \left(\frac{\mathbf{\Theta}_v^{*T} \nabla \ell_i^{(j)}(\boldsymbol{\beta}^*)}{(\Theta_{vv}^*)^{1/2}} \overline{\Delta}^{(j)} + \big(\widehat{\mathbf{\Theta}}_v^{(j)} - \mathbf{\Theta}_v^*\big)^T \frac{\nabla \ell_i(\boldsymbol{\beta}^*)}{(\mathbf{\Theta}_{vv}^*)^{1/2}} \overline{\Delta}^{(j)}\right) \\
&= \sum_{j=1}^k \sum_{i\in\mathcal{I}_j} \big(\Delta_{21,i}^{(j)} + \Delta_{22,i}^{(j)}\big),
\end{aligned}$$

where $\left|\sum_{j=1}^k \sum_{i\in\mathcal{I}_j} \Delta_{21,i}^{(i)}\right| \leq \left|\sum_{j=1}^k \sum_{i\in\mathcal{I}_j} \xi_{i,v}^{(j)}\right| \left|\overline{\Theta}_{vv}^{1/2} - (\Theta_{vv}^*)^{1/2}\right|$. Since $\Theta_{vv}^* \geq 0$, $\overline{\Theta}_{vv}^{1/2} = |\overline{\Theta}_{vv}|^{1/2} = |\overline{\Theta}_{vv} - \Theta_{vv}^* + \Theta_{vv}^*|^{1/2} \leq |\overline{\Theta}_{vv} - \Theta_{vv}^*|^{1/2} + (\Theta_{vv}^*)^{1/2}$. Similarly

$$(\Theta_{vv}^*)^{1/2} = |\Theta_{vv}^*|^{1/2} = |\Theta_{vv}^* - \overline{\Theta}_{vv} + \overline{\Theta}_{vv}|^{1/2} \leq |\Theta_{vv}^* - \overline{\Theta}_{vv}|^{1/2} + \overline{\Theta}_{vv}^{1/2},$$

yielding $|\overline{\Theta}_{vv}^{1/2} - (\Theta_{vv}^*)^{1/2}| \leq |\overline{\Theta}_{vv} - \Theta_{vv}^*|^{1/2}$ and consequently, by assumption (A5),

$$\left|\overline{\Delta}^{(j)}\right| = \left|\overline{\Theta}_{vv}^{1/2} - (\Theta_{vv}^*)^{1/2}\right| = o_{\mathbb{P}}(1).$$

Invoking (A9) and the Lindeberg-Feller CLT, $\left|\sum_{j=1}^k \sum_{i\in\mathcal{I}_j} \Delta_{21,i}^{(i)}\right| = o_{\mathbb{P}}(1)$. Similarly

$$\left|\sum_{j=1}^k \sum_{i\in\mathcal{I}_j} \Delta_{22,i}^{(j)}\right| \leq \max_{1\leq j\leq k} \|\widehat{\mathbf{\Theta}}_v^{(j)} - \mathbf{\Theta}_v^*\|_1 \left|\overline{\Delta}^{(j)}\right| \left|\left(\mathbf{\Theta}_v^{*T}\mathbf{\Theta}_v^*\right)^{-1/2} \sum_{j=1}^k \sum_{i\in\mathcal{I}_j} \xi_{iv}^{(j)}\right| = o_{\mathbb{P}}(1).$$

Combining all terms in the decomposition (E.7) delivers the result. □

(B1)-(B5) of Condition E.9 are used in the proofs of subsequent lemmas.

Condition E.9. (B1) $\|\boldsymbol{w}^*\|_1 \lesssim s_1$, $\|J^*\|_{\max} < \infty$ and for any $\delta \in (0,1)$,

$$\mathbb{P}\Big(\|\widehat{\boldsymbol{\beta}}_{-v}^{\lambda} - \boldsymbol{\beta}_{-v}^*\|_1 \gtrsim n^{-1/2} s\sqrt{\log(d/\delta)}\Big) < \delta$$

and

$$\mathbb{P}\Big(\|\widehat{\boldsymbol{w}} - \boldsymbol{w}^*\|_1 \gtrsim n^{-1/2} s_1\sqrt{\log(d/\delta)}\Big) < \delta.$$

(B2) For any $\delta \in (0,1)$,

$$\mathbb{P}\Big(\|\nabla_{-v}\ell_n(\beta_v^*, \boldsymbol{\beta}_{-v}^*)\|_\infty \gtrsim n^{-1/2}\sqrt{\log(d/\delta)}\Big) < \delta.$$

(B3) Suppose $\widehat{\boldsymbol{\beta}}_{-v}^{\lambda}$ satisfies (B1). Define

$$H_v := \big(\nabla_{v,-v}^2 \ell_n(\beta_v^*, \boldsymbol{\beta}_{-v,\alpha}) - \widehat{\boldsymbol{w}}^T \nabla_{-v,-v}^2 \ell_n(\beta_v^*, \boldsymbol{\beta}_{-v,\alpha})\big) \cdot (\widehat{\boldsymbol{\beta}}_{-v}^{\lambda} - \boldsymbol{\beta}_{-v}^*).$$

Then for $\boldsymbol{\beta}_{-v,\alpha} = \alpha\boldsymbol{\beta}_{-v}^* + (1-\alpha)\widehat{\boldsymbol{\beta}}_{-v}^{\lambda}$ and for any $\delta \in (0,1)$,

$$\mathbb{P}\left(\sup_{\alpha\in[0,1]} |H_v| \gtrsim s_1 s \frac{\log(d/\delta)}{n}\right) < \delta.$$

(B4) There exists a constant $C > 0$ such that $C < I^*_{\theta|\boldsymbol{\gamma}} < \infty$, and for $\boldsymbol{v}^* = (1, -\boldsymbol{w}^{*T})^T$, it holds that

$$\frac{\sqrt{n}\boldsymbol{v}^{*T}\nabla\ell_n(\beta_v^*, \boldsymbol{\beta}^*_{-v})}{\sqrt{\boldsymbol{v}^{*T}\boldsymbol{J}^*\boldsymbol{v}^*}} \rightsquigarrow N(0,1).$$

(B5) For any δ, if there exists an estimator $\widetilde{\boldsymbol{\beta}} = (\widetilde{\beta}_v^T, \widetilde{\boldsymbol{\beta}}_{-v}^T)^T$ satisfying $\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \leq Cs\sqrt{n^{-1}\log(d/\delta)}$ with probability $> 1-\delta$, then

$$\mathbb{P}\Big(\big\|\nabla^2\ell_n(\widetilde{\boldsymbol{\beta}}) - J^*\big\|_{\max} \gtrsim n^{-1/2}\sqrt{\log(d/\delta)}\Big) < \delta.$$

The proof of Theorem 3.11 is an application of Lemma E.13. To apply this Lemma, we must first verify (B1) to (B4) of Condition E.9. We do this in Lemma E.10.

Lemma E.10. Under the requirements of Theorem 3.11, (B1) - (B4) of Condition E.9 are fulfilled.

Proof. **Verification of (B1)**. As stated in Theorem 3.11, $\|\boldsymbol{w}^*\|_1 = O(s_1)$ and $\|J^*\|_{\max} < \infty$ by part (i) of Condition 3.6. The rest of (B1) follows from the proof of Lemma C.3 of Ning and Liu (2014).

Verification of (B2). Let $\boldsymbol{X}_i = (Q_i, \boldsymbol{Z}_i^T)^T$. Since $\|\nabla_{\boldsymbol{\gamma}}\ell_n(\boldsymbol{\beta}^*)\|_\infty = \big\|-\frac{1}{n}\sum_{i=1}^n \big(Y_i - b'(\boldsymbol{X}_i^T\boldsymbol{\beta}^*)\big)\boldsymbol{Z}_i\big\|_\infty$, since the product of a sub-Gaussian random variable and a bounded random variable is sub-Gaussian, and since $\mathbb{E}[\nabla_{\boldsymbol{\gamma}}\ell_n(\boldsymbol{\beta}^*)] = 0$, we have by Condition 3.6, Bernstein's inequality and the union bound

$$\mathbb{P}\big(\|\nabla_{\boldsymbol{\gamma}}\ell_n(\boldsymbol{\beta}^*)\|_\infty > t\big) < (d-1)\exp\{-nt^2/M^2\sigma_b^2\}.$$

Setting $2(d-1)\exp\{-nt^2/M^2\sigma_b^2\} = \delta$ and solving for t delivers the result.

Verification of (B3) Let $\boldsymbol{\beta}^*_\alpha = (\theta^*, \boldsymbol{\gamma}_\alpha)$ and decompose the object of interest as

(E.8)

$$\big|\big(\nabla^2_{v,-v}\ell_n(\beta_v^*, \boldsymbol{\beta}_{-v,\alpha}) - \widehat{\boldsymbol{w}}^T\nabla^2_{-v,-v}\ell_n(\beta_v^*, \boldsymbol{\beta}_{-v,\alpha})\big)(\widehat{\boldsymbol{\beta}}^\lambda_{-v} - \boldsymbol{\beta}^*_{-v})\big| \leq \sum_{t=1}^5 \big|\Delta_t\big|,$$

where the terms Δ_1 - Δ_5 are given by $\Delta_1 = \nabla^2_{v,-v}\ell_n(\boldsymbol{\beta}^*_\alpha) - \nabla^2_{v,-v}\ell_n(\boldsymbol{\beta}^*)$,

$$\Delta_2 = \nabla^2_{v,-v}\ell_n(\boldsymbol{\beta}^*) - \boldsymbol{w}^{*T}J^*_{-v,-v},\ \Delta_4 = \boldsymbol{w}^{*T}\big(\nabla^2_{-v,-v}\ell_n(\boldsymbol{\beta}^*) - \nabla^2_{-v,-v}\ell_n(\boldsymbol{\beta}^*_\alpha)\big),$$
$$\Delta_3 = \boldsymbol{w}^{*T}\big(J^*_{-v,-v} - \nabla^2_{-v,-v}\ell_n(\boldsymbol{\beta}^*)\big),\ \Delta_5 = (\boldsymbol{w}^{*T} - \widehat{\boldsymbol{w}}^T)\nabla^2_{-v,-v}\ell_n(\boldsymbol{\beta}^*_\alpha).$$

We have the following bounds

$$\begin{aligned}|\Delta_1| &= \Big|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{Z}_i\boldsymbol{Z}_i^T(\widehat{\boldsymbol{\beta}}^{\lambda}_{-v}-\boldsymbol{\beta}^*_{-v})\big(\ell_i''(\boldsymbol{X}_i^T\boldsymbol{\beta}^*_{\alpha})-\ell_i''(\boldsymbol{X}_i^T\boldsymbol{\beta}^*)\big)\Big|\\ &\le \max_{1\le i\le n}K_i\max_{1\le i\le n}\|\boldsymbol{X}_i\|_\infty\Big\|\frac{1}{n}\boldsymbol{Z}(\widehat{\boldsymbol{\beta}}_{-v}-\boldsymbol{\beta}^*_{-v})\Big\|_2^2,\end{aligned}$$

$$|\Delta_2|\le\big\|\nabla^2_{v,-v}\ell_n(\boldsymbol{\beta}^*)-J^*_{v,-v}\big\|_\infty\|\widehat{\boldsymbol{\beta}}^{\lambda}_{-v}-\boldsymbol{\beta}^*_{-v}\|_1,$$
$$|\Delta_3|\le\|\boldsymbol{w}\|_1\big\|J^*_{-v,-v}-\nabla^2_{-v,-v}\ell_n(\boldsymbol{\beta}^*)\big\|_{\max}\|\widehat{\boldsymbol{\beta}}^{\lambda}_{-v}-\boldsymbol{\beta}^*_{v}\|_1,$$

$$\begin{aligned}|\Delta_4| &= \big|\boldsymbol{w}^{*T}\big(\nabla^2_{-v,-v}\ell_n(\boldsymbol{\beta}^*)-\nabla^2_{-v,-v}\ell_n(\boldsymbol{\beta}^*_v)\big)(\widehat{\boldsymbol{\gamma}}^{\lambda}-\boldsymbol{\lambda}^*)\big|\\ &\le \max_{1\le i\le n}K_i\|\boldsymbol{w}^*\|_1\Big\|\frac{1}{n}\boldsymbol{Z}(\widehat{\boldsymbol{\beta}}^{\lambda}_{-v}-\boldsymbol{\beta}^*_{-v})\Big\|_2^2,\end{aligned}$$

and $|\Delta_5|\le\|\boldsymbol{w}^*-\widehat{\boldsymbol{w}}\|_1\big\|\nabla_{-v,-v}\ell_n(\boldsymbol{\beta}^*_v)\big\|_{\max}\|\widehat{\boldsymbol{\beta}}^{\lambda}_{-v}-\boldsymbol{\beta}^*_{-v}\|_1$. Let $\varepsilon=\delta/5$. Then by Condition 3.6 and Lemma E.4

$$\mathbb{P}\Big(|\Delta_1|\gtrsim s\frac{\log(d/\varepsilon)}{n}\Big)<\varepsilon\quad\text{and}\quad\mathbb{P}\Big(|\Delta_4|\gtrsim ss_1\frac{\log(d/\varepsilon)}{n}\Big)<\varepsilon.$$

Noting the $\boldsymbol{\beta}^*$ itself satisfies the requirements on $\widetilde{\boldsymbol{\beta}}$ in (B5), Lemma E.11 and Condition 2.1 together give

$$\mathbb{P}\Big(|\Delta_2|\gtrsim s_1\frac{\log(d/\varepsilon)}{n}\Big)<\varepsilon\quad\text{and}\quad\mathbb{P}\Big(|\Delta_3|\gtrsim s_1s\frac{\log(d/\varepsilon)}{n}\Big)<\varepsilon.$$

By (B1) verified above and noting that

$$\begin{aligned}\big\|\nabla_{-v,-v}\ell_n(\boldsymbol{\beta}^*_v)\big\|_{\max} &\le \big\|\nabla_{-v,-v}\ell_n(\boldsymbol{\beta}^*_v)-\nabla_{-v,-v}\ell_n(\boldsymbol{\beta}^*)\big\|_{\max}\\ &\quad+\big\|\nabla_{-v,-v}\ell_n(\boldsymbol{\beta}^*)\big\|_{\max},\end{aligned}$$

the proof of Lemma E.11 delivers $\mathbb{P}\Big(|\Delta_5|\gtrsim s_1s\log(d/\varepsilon)/n\Big)<\varepsilon$. Combining the bounds, we finally have

$$\mathbb{P}\Big(\sup_{\alpha\in[0,1]}H_v\gtrsim s_1s\frac{\log(d/\delta)}{n}\Big)<\delta.$$

Verification of (B4). See Ning and Liu (2014), proof of Lemma C.2. □

In the following lemma, we verify (B5) under the same conditions.

Lemma E.11. Under Conditions 3.6 and 2.1, (B5) of Condition E.9 is fulfilled.

PROOF. We obtain a tail probability bound for Δ_1 and Δ_2 in the decomposition

$$\begin{aligned}\|\nabla^2 \ell_n(\widetilde{\boldsymbol{\beta}}) - J^*\|_{\max} &\leq \|\nabla^2 \ell_n(\widetilde{\boldsymbol{\beta}}) - \nabla^2 \ell_n(\boldsymbol{\beta}^*)\|_{\max} + \|\nabla^2 \ell_n(\boldsymbol{\beta}^*) - J^*\|_{\max} \\ &= \Delta_1 + \Delta_2.\end{aligned}$$

For the control over Δ_1, note that by Condition 3.6 (ii) and (iii),

$$\big|[\nabla^2 \ell_n(\boldsymbol{\beta}^*)]_{jk}\big| \leq \big|b''(\boldsymbol{X}_i^T \boldsymbol{\beta}^*)\big|\big|X_{ij} X_{ik}\big| \leq U_2 M^2.$$

Hence Hoeffding's inequality and the union bound deliver

$$\mathbb{P}(\Delta_2 > t) = \mathbb{P}\Big(\|\nabla^2 \ell_n(\boldsymbol{\beta}^*) - J^*\|_{\max} > t\Big) \leq 2d^2 \exp\Big\{-\frac{nt^2}{8U_2^2 M^4}\Big\}. \tag{E.9}$$

For the control over Δ_1, we have by Lemma E.5,

$$\begin{aligned}\big|[\nabla^2 \ell_n(\widetilde{\boldsymbol{\beta}}) - \nabla^2 \ell_n(\boldsymbol{\beta}^*)]_{jk}\big| &= \big|\big(b''(\boldsymbol{X}_i^T \widetilde{\boldsymbol{\beta}}) - b''(\boldsymbol{X}_i^T \boldsymbol{\beta}^*)\big) X_{ij} X_{ik}\big| \\ &\leq M^3 U_3 \|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \leq M^3 U_3 s \sqrt{n^{-1} \log(d/\delta)}\end{aligned}$$

with probability $> 1 - \delta$. Hoeffding's inequality and the union bound again deliver

$$\begin{aligned}\mathbb{P}(\Delta_1 > t) &= \mathbb{P}\Big(\|\nabla^2_{\eta\eta} \ell_n(\widetilde{\boldsymbol{\beta}}) - \nabla^2_{\eta\eta} \ell_n(\boldsymbol{\beta}^*)\|_{\max} > t\Big) \\ &\leq 2d^2 \exp\Big\{-\frac{n^2 t^2}{8U_3^2 M^6 s^2 \log(d/\delta)}\Big\}.\end{aligned} \tag{E.10}$$

Combining the bounds from equations (E.9) and (E.10) we have

$$\begin{aligned}&\mathbb{P}\Big(\|\nabla^2 \ell(\widetilde{\boldsymbol{\beta}}) - J^*\|_{\max} > t\Big) \\ &\leq 2d^2 \Big(\exp\Big\{-\frac{nt^2}{8U_3^2 M^4}\Big\} + \exp\Big\{-\frac{n^2 t^2}{8U_3^2 M^6 s^2 \log(d/\delta)}\Big\}\Big).\end{aligned}$$

Setting each term equal to $\delta/2$, solving for t and ignoring the relative magnitude of constants, we have $t = U_3 \max\{n^{-1} s \log(d/\delta), n^{-1/2}\sqrt{\log(d/\delta)}\} = U_3 n^{-1/2} \log(d/\delta)$, thus verifying (B5). □

Lemma E.12. For each $j \in \{1, \dots, k\}$, let $\boldsymbol{\beta}_{-v,\alpha_j} = \alpha_j \widehat{\boldsymbol{\beta}}^{\lambda}_{-v}(\mathcal{D}_j) + (1 - \alpha_j)\boldsymbol{\beta}^*_{-v}$, for some $\alpha_j \in [0,1]$, where $\widehat{\boldsymbol{\beta}}^{\lambda}_{-v}(\mathcal{D}_j)$ is defined in equation (2.2). Define

$$\begin{aligned} \Delta_1^{(j)} &= (\widehat{\boldsymbol{w}}(\mathcal{D}_j) - \boldsymbol{w}^*)^T \nabla_{-v} \ell^{(j)}_{n_k}(\beta^*_v, \boldsymbol{\beta}^*_{-v}) \quad \text{and} \\ \Delta_2^{(j)} &= \big(\nabla^2_{v,-v} \ell^{(j)}_{n_k}(\beta^*_v, \boldsymbol{\beta}_{-v,\alpha_j}) - \widehat{\boldsymbol{w}}^T \nabla_{-v,-v} \ell^{(j)}_{n_k}(\beta^*_v, \boldsymbol{\beta}_{-v,\alpha_j})\big)(\widehat{\boldsymbol{\beta}}^{\lambda}_{-v} - \boldsymbol{\beta}^*_{-v}). \end{aligned}$$

Under (B1) - (B3) of Condition E.9, $\Big|k^{-1}\sum_{j=1}^k \Delta_1^{(j)}\Big| = o_{\mathbb{P}}\big(n^{-1/2}\big)$ and $\Big|k^{-1}\sum_{j=1}^k \Delta_2^{(j)}\Big| = o_{\mathbb{P}}\big(n^{-1/2}\big)$ whenever $k \ll d$ is chosen to satisfy $k = o\big((s_1 \log d)^{-1}\sqrt{n}\big)$.

Proof. By Hölder's inequality,

$$\begin{aligned} \big|\Delta_1^{(j)}\big| &= \big|\big(\boldsymbol{w}^* - \widehat{\boldsymbol{w}}(\mathcal{D}_j)\big)^T \nabla_{-v} \ell^{(j)}_{n_k}(\beta^*_v, \boldsymbol{\beta}^*_{-v})\big| \\ &\le \|\widehat{\boldsymbol{w}}(\mathcal{D}_j) - \boldsymbol{w}^*\|_1 \|\nabla_{-v} \ell^{(j)}_{n_k}(\beta^*_v, \boldsymbol{\beta}^*_{-v})\|_\infty, \end{aligned}$$

hence, for any t,

$$\big\{\big|\Delta_1^{(j)}\big| > t\big\} \subseteq \big\{\|\widehat{\boldsymbol{w}}(\mathcal{D}_j) - \boldsymbol{w}^*\|_1 \|\nabla_{-v} \ell^{(j)}_{n_k}(\beta^*_v, \boldsymbol{\beta}^*_{-v})\|_\infty > t\big\}.$$

Take $t = vq$ where $v = Cn^{-1/2}s_1\sqrt{k\log(d/\delta)}$ and $q = Cn^{-1/2}\sqrt{k\log(d/\delta)}$. Define two events E_1 and E_2 as following.

$$\begin{aligned} E_1 &:= \big\{\|\widehat{\boldsymbol{w}}(\mathcal{D}_j) - \boldsymbol{w}^*\|_1 \|\nabla_{-v} \ell^{(j)}_{n_k}(\beta^*_v, \boldsymbol{\beta}^*_{-v})\|_\infty > vq\big\}, \\ E_2 &:= \Big\{\frac{\|\widehat{\boldsymbol{w}}(\mathcal{D}_j) - \boldsymbol{w}^*\|_1}{v} \le 1\Big\}. \end{aligned}$$

Then we obtain that $\mathbb{P}(E_1) = \mathbb{P}(E_1 \cap E_2) + \mathbb{P}(E_1 \cap E_2^c)$, which is not greater than 2δ by (B1) and (B2) of Condition E.9. Hence the union bound delivers

$$\begin{aligned} \mathbb{P}\Big(\big|\sum_{j=1}^k \Delta_1^{(j)}\big| > kvq\Big) &\le \mathbb{P}\Big(\cup_{j=1}^k\big\{\big|\Delta_1^{(j)}\big| > vq\big\}\Big) \le \sum_{j=1}^k \mathbb{P}\Big(\big|\Delta_1^{(j)}\big| > vq\Big) \\ &\le 2k\delta = o(1) \end{aligned}$$

for $\delta = o(k^{-1})$. Taking $\delta = k^{-1}$ for $\alpha > 0$ arbitrarily small in the definition of v and q, the requirement is $ks_1 \log d = o\big(\sqrt{n}\big)$ and $ks_1 \log k = o\big(\sqrt{n}\big)$ for $\alpha > 0$ arbitrarily small. Since $k \ll d$, $k^{-1}\sum_{j=1}^k \Delta_1^{(j)} = o_{\mathbb{P}}\big(n^{-1/2}\big)$ with $k = o\big((s_1 \log d)^{-1}\sqrt{n}\big)$. Next, consider $|\Delta_2^{(j)}| \le \sup_{\alpha\in[0,1]} |G_v|$, where

$$G_v := (\nabla^2_{v,-v} \ell^{(j)}_{n_k}(\beta^*_v, \boldsymbol{\beta}_{-v,\alpha}) - \widehat{\boldsymbol{w}}^T \nabla^2_{-v,-v} \ell^{(j)}_{n_k}(\beta^*_v, \boldsymbol{\beta}_{-v,\alpha})\big)(\widehat{\boldsymbol{\beta}}^{\lambda}_{-v}(\mathcal{D}_j) - \boldsymbol{\beta}^*_{-v}).$$

By (B3) of Condition E.9, $\mathbb{P}\big(\big|\Delta_2^{(j)}\big| \geq t\big) < \delta$ for $t \asymp s_1 s n^{-1} k \log(d/\delta)$, hence, proceeding in an analogous fashion to in the control over $k^{-1}\sum_{j=1}^{k}\Delta_1^{(j)}$, we obtain

$$\mathbb{P}\Big(\Big|\sum_{j=1}^{k}\Delta_2^{(j)}\Big| > kt\Big) \leq \mathbb{P}\left(\cup_{j=1}^{k}|\Delta_2^{(j)}| > t\right) \leq \sum_{j=1}^{k}\mathbb{P}\left(|\Delta_2^{(j)}| > t\right) \leq k\delta = o(1)$$

for $\delta = o(k^{-1})$. Hence $k^{-1}\sum_{j=1}^{k}\Delta_2^{(j)} = o_{\mathbb{P}}\big(n^{-1/2}\big)$ with $k = o\big((s_1 s \log d)^{-1} \cdot n^{3/2}\big)$. Since $(s_1 \log d)^{-1}\sqrt{n} = o\big((s_1 s \log d)^{-1} n^{3/2}\big)$, $k^{-1}\sum_{j=1}^{k}\big(\Delta_1^{(j)} + \Delta_2^{(j)}\big) = o_{\mathbb{P}}\big(n^{-1/2}\big)$ requires $k = o\big((s_1 \log d)^{-1}\sqrt{n}\big)$. □

Lemma E.13. Under (B1) - (B4) of Condition E.9, with $k \ll d$ chosen to satisfy the scaling $k = o\big(((s \vee s_1)\log d)^{-1}\sqrt{n}\big)$,

$$\frac{1}{k}\sum_{j=1}^{k}\widehat{S}^{(j)}(\beta_v^*, \widehat{\boldsymbol{\gamma}}^{\lambda}(\mathcal{D}_\mathrm{j})) = \frac{1}{k}\sum_{j=1}^{k} S^{(j)}(\beta_v^*, \boldsymbol{\beta}_{-v}^*) + o_{\mathbb{P}}(n^{-1/2}) \;\text{ and}$$

$$\lim_{n\to\infty}\sup_t |\mathbb{P}((J_{v|-v}^*)^{-1/2}\sqrt{n}\frac{1}{k}\sum_{j=1}^{k} S^{(j)}(\beta_v^*, \boldsymbol{\beta}_{-v}^*) < t) - \Phi(t)| \to 0.$$

Proof. Recall

$$S^{(j)}(\beta_v^*, \boldsymbol{\beta}_{-v}^*) = \nabla_v \ell_{n_k}^{(j)}(\beta_v^*, \boldsymbol{\beta}_{-v}^*) - \boldsymbol{w}^{*T}\nabla_{-v}\ell_{n_k}^{(j)}(\beta_v^*, \boldsymbol{\beta}_{-v}^*).$$

Through a mean value expansion of $\widehat{S}^{(j)}(\beta_v^*, \widehat{\boldsymbol{\beta}}_{-v}^{\lambda}(\mathcal{D}_j))$ around $\boldsymbol{\beta}_{-v}^*$, we have for each $j \in \{1, \ldots, k\}$,

$$\begin{aligned}\widehat{S}^{(j)}\big(\beta_v^*, \widehat{\boldsymbol{\beta}}_{-v}^{\lambda}(\mathcal{D}_j)\big) &= \nabla_v \ell_{n_k}^{(j)}\big(\beta_v^*, \widehat{\boldsymbol{\beta}}_{-v}^{\lambda}(\mathcal{D}_j)\big) - \widehat{\boldsymbol{w}}(\mathcal{D}_j)^T\nabla_{-v}\ell_{n_k}^{(j)}\big(\beta_v^*, \widehat{\boldsymbol{\beta}}_{-v}^{\lambda}(\mathcal{D}_j)\big)\\ &= S^{(j)}(\beta_v^*, \boldsymbol{\beta}_{-v}^*) + \Delta_1^{(j)} + \Delta_2^{(j)},\end{aligned}$$

for some $\boldsymbol{\beta}_{-v,\alpha} = \alpha\widehat{\boldsymbol{\beta}}_{-v}(\mathcal{D}_j) + (1-\alpha)\boldsymbol{\beta}_{-v}^*$, where

$$\begin{aligned}\Delta_1^{(j)} &= \big(\boldsymbol{w}^* - \widehat{\boldsymbol{w}}(\mathcal{D}_j)\big)^T\nabla_{-v}\ell_{n_k}^{(j)}(\beta_v^*, \boldsymbol{\beta}_{-v}^*)\\ \Delta_2^{(j)} &= \Big[\nabla_{v,-v}^2\ell_{n_k}^{(j)}(\beta_v^*, \boldsymbol{\beta}_{-v,\alpha}) - \widehat{\boldsymbol{w}}(\mathcal{D}_j)^T\nabla_{-v,-v}^2\ell_{n_k}^{(j)}(\beta_v^*, \boldsymbol{\beta}_{-v,\alpha})\Big](\widehat{\boldsymbol{\beta}}_{-v}^{\lambda}(\mathcal{D}_j) - \boldsymbol{\beta}_{-v}^*).\end{aligned}$$

Here $\boldsymbol{h}_v = \nabla_{v,-v}^2\ell_{n_k}^{(j)}(\beta_v^*, \boldsymbol{\beta}_{-v,\alpha}) - \widehat{\boldsymbol{w}}(\mathcal{D}_j)^T\nabla_{-v,-v}^2\ell_{n_k}^{(j)}(\beta_v^*, \boldsymbol{\beta}_{-v,\alpha})$. It follows that

(E.11)

$$\begin{aligned}\frac{1}{k}\sum_{j=1}^{k}\widehat{S}^{(j)}\big(\beta_v^*,\widehat{\boldsymbol{\beta}}_{-v}^{\lambda}(\mathcal{D}_j)\big) &= \frac{1}{k}\sum_{j=1}^{k}S^{(j)}(\beta_v^*,\boldsymbol{\beta}_{-v}^*)+\frac{1}{k}\sum_{j=1}^{k}\big(\Delta_1^{(j)}+\Delta_2^{(j)}\big)\\ &= \frac{1}{k}\sum_{j=1}^{k}S^{(j)}(\theta^*,\boldsymbol{\gamma}^*)+o_{\mathbb{P}}(n^{-1/2})\end{aligned}$$

by Lemma E.12 whenever $k=o\big((s_1\log d)^{-1}\sqrt{n}\big)$. Observe

$$\sqrt{n}\Big(k^{-1}\sum_{j=1}^{k}S^{(j)}(\beta_v^*,\boldsymbol{\beta}_{-v}^*)\Big)=\sqrt{n}(1,-\boldsymbol{w}^{*T})\Big(\frac{1}{k}\sum_{j=1}^{k}\nabla\ell_{n_k}^{(j)}(\beta_v^*,\boldsymbol{\beta}_{-v}^*)\Big)\quad\text{and}$$

$$J_{v|-v}^*=(1,-\boldsymbol{w}^{*T})J^*(1,-\boldsymbol{w}^{*T})^T.$$

So $\sqrt{n}\frac{1}{k}\sum_{j=1}^{k}S^{(j)}(\beta_v^*,\boldsymbol{\beta}_{-v}^*)\rightsquigarrow N(0,J_{v|-v}^*)$ by Condition (B4). Similar to Corollary 3.9, we apply the Berry-Essen inequality to show that

$$\sup_t|\mathbb{P}(\sqrt{n}\frac{1}{k}\sum_{j=1}^{k}S^{(j)}(\beta_v^*,\boldsymbol{\beta}_{-v}^*)<t)-\Phi(t)|\to 0.$$

□

Lemma E.14. Under Condition (B1), for any $\delta\in(0,1)$,

$$\mathbb{P}\Big(\|\overline{\boldsymbol{w}}-\boldsymbol{w}^*\|_1>Cn^{-1/2}s_1\sqrt{k\log(d/\delta)}\Big)<k\delta,$$
$$\mathbb{P}\Big(\|\overline{\boldsymbol{\beta}}_{-v}-\boldsymbol{\beta}_{-v}^*\|_1>Cn^{-1/2}s\sqrt{k\log(d/\delta)}\Big)<k\delta.$$

PROOF. Set $t=Cs_1\sqrt{n^{-1}(k\log(d/\delta))}$ and note

$$\mathbb{P}\big(\|\sum_{j=1}^{k}(\widehat{\boldsymbol{w}}(\mathcal{D}_j)-\boldsymbol{w}^*)\|_1>kt\big)\le\sum_{j=1}^{k}\mathbb{P}\big(\|\overline{\boldsymbol{w}}-\boldsymbol{w}^*\|_1>t\big).$$

by the union bound. Then by Condition (B1),

$$\mathbb{P}\Big(\|\overline{\boldsymbol{w}}-\boldsymbol{w}^*\|_1>Cn^{-1/2}s_1\sqrt{k\log(d/\delta)}\Big)<k\delta.$$

The proof of the second bound is analogous, setting $t=Cs\sqrt{n^{-1}(k\log(d/\delta))}$.

□

Lemma E.15. Suppose (B5) of Condition E.9 is satisfied. For any δ, if there exists an estimator $\widetilde{\boldsymbol{\beta}} = (\widetilde{\beta}_v^T, \widetilde{\boldsymbol{\beta}}_{-v}^T)^T$ satisfying $\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \le Cs\sqrt{n^{-1}\log(d/\delta)}$ with probability $1-\delta$, then

$$\mathbb{P}\Big(\Big\|\frac{1}{k}\sum_{j=1}^{k}\nabla^2\ell_{n_k}^{(j)}(\widetilde{\boldsymbol{\beta}}) - J^*\Big\|_{\max} > Cn^{-1/2}\sqrt{k\log(d/\delta)}\Big) < k\delta.$$

Proof. The proof follows from (B5) in Condition E.9 via an analogous argument to that of Lemma E.14, taking $t = C\sqrt{n^{-1}(k\log(d/\delta))}$. □

Lemma E.16. Suppose (B1)-(B5) of Condition E.9 are fulfilled. Then for any $k \ll d$ satisfying $k = o\big(((s \vee s_1)\log d)^{-1}\sqrt{n}\big)$, $|\overline{J}_{\theta|\boldsymbol{\gamma}} - J^*_{v|-v}| = o_{\mathbb{P}}(1)$.

Proof. Recall that $J^*_{v|-v} = J^*_{v,v} - \boldsymbol{J}^*_{v,-v}J^{*-1}_{-v,-v}\boldsymbol{J}^*_{-v,v}$ and

$$\overline{J}_{v|-v} = \frac{1}{k}\sum_{j=1}^{k}\big(\nabla_{v,v}\ell_{n_k}^{(j)}(\overline{\beta}_v^d, \overline{\boldsymbol{\beta}}_{-v}) - \overline{w}^T\nabla^2_{-v,v}\ell_{n_k}^{(j)}(\overline{\beta}_v^d, \overline{\boldsymbol{\beta}}_{-v}\big),$$

so $\big|\overline{J}_{v|-v} - J^*_{v|-v}\big| = \delta_1 + \delta_2$, where

$$\Delta_1 = \Big|\frac{1}{k}\sum_{j=1}^{k}\nabla_{v,v}\ell_{n_k}^{(j)}(\overline{\beta}_v^d, \overline{\boldsymbol{\beta}}_{-v}) - J^*_{v,v}\Big| \quad \text{and}$$

$$\Delta_2 = \Big|\overline{\boldsymbol{w}}^T\big(\frac{1}{k}\sum_{j=1}^{k}\nabla^2_{-v,v}\ell_{n_k}^{(j)}(\overline{\beta}_v^d, \overline{\boldsymbol{\beta}}_{-v}) - \boldsymbol{w}^{*T}\boldsymbol{J}^*_{-v,v}\big)\Big|.$$

Let $\widetilde{\boldsymbol{\beta}} = (\overline{\beta}_v^d, \overline{\boldsymbol{\beta}}_{-v})$ and note that $\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1$ satisfies the clause in (B5) of Condition E.9 by Lemma E.14 when $k = o\big(((s \vee s_1)\log d)^{-1}\sqrt{n}\big)$. Hence $\Delta_1 = o_{\mathbb{P}}(1)$ by Lemma E.15.

$$\Delta_2 \le \underbrace{\Big|(\overline{\boldsymbol{w}} - \boldsymbol{w}^*)^T\Big(\frac{1}{k}\sum_{j=1}^{k}\nabla^2_{-v,v}\ell_{n_k}^{(j)}(\overline{\beta}_v^d, \overline{\boldsymbol{\beta}}_{-v}) - \boldsymbol{J}^*_{-v,v}\Big)\Big|}_{\Delta_{21}}$$

$$+ \underbrace{\big|(\overline{\boldsymbol{w}} - \boldsymbol{w}^*)^T\boldsymbol{J}^*_{-v,v}\big|}_{\Delta_{22}} + \underbrace{\Big|\boldsymbol{w}^{*T}\Big(\frac{1}{k}\sum_{j=1}^{k}\nabla^2_{-v,v}\ell_{n_k}^{(j)}(\overline{\beta}_v^d, \overline{\boldsymbol{\beta}}_{-v}) - \boldsymbol{J}^*_{-v,v}\Big)\Big|}_{\Delta_{23}}.$$

By the fact that $\|J^*\|_{\max} < \infty$ and $\|\boldsymbol{w}^*\|_1 \le Cs_1$ by (B1) of Condition E.9, an application of Lemmas E.14 and E.15 delivers

$$\begin{aligned}
\Delta_{21} &\le \|\overline{\boldsymbol{w}} - \boldsymbol{w}^*\|_1 \Big\| \frac{1}{k}\sum_{j=1}^{k} \nabla^2_{-v,v}\ell^{(j)}_{n_k}(\overline{\beta}^d_v, \overline{\boldsymbol{\beta}}_{-v}) - \boldsymbol{J}^*_{-v,v}\Big\|_\infty = o_{\mathbb{P}}(1), \\
\Delta_{22} &\le \|\overline{\boldsymbol{w}} - \boldsymbol{w}^*\|_1 \|\boldsymbol{J}^*_{-v,v}\|_\infty = o_{\mathbb{P}}(1), \\
\Delta_{23} &\le \Big\| \frac{1}{k}\sum_{j=1}^{k} \nabla^2_{-v,v}\ell^{(j)}_{n_k}(\overline{\beta}^d_v, \overline{\boldsymbol{\beta}}_{-v}) - \boldsymbol{J}^*_{-v,v}\Big\|_\infty \|\boldsymbol{w}^*\|_1 = o_{\mathbb{P}}(1)
\end{aligned}$$

for $k = o\big((s_1 \log d)^{-1} n\big)$, a fortiori for $k = o\big(((s \vee s_1)\log d)^{-1}\sqrt{n}\big)$. Hence $\big|\overline{J}_{v|-v} - J^*_{v|-v}\big| = o_{\mathbb{P}}(1)$. □

APPENDIX F: AUXILIARY LEMMAS FOR ESTIMATION

In this section, we provide the proofs of the technical lemmas and theorems for the divide and conquer estimation. Using Lemma 7.1 we can derive Lemma F.1, which serves as a crucial step in establishing Lemma 4.1.

Lemma F.1. Suppose Conditions 3.1 and 3.2 are fulfilled. Let $\lambda \asymp \sqrt{k \log d/n}$ and $\vartheta_1 \asymp \sqrt{k \log d/n}$. With $k = o((s\log d)^{-1}\sqrt{n})$, $\sqrt{n}(\overline{\boldsymbol{\beta}}^d - \boldsymbol{\beta}^*) = \boldsymbol{Z} + \boldsymbol{\Delta}$, where $\boldsymbol{Z} = \frac{1}{\sqrt{k}}\sum_{j=1}^k \frac{1}{\sqrt{n_k}} M^{(j)} X^{(j)T} \boldsymbol{\varepsilon}^{(j)}$ and $\|\boldsymbol{\Delta}\|_\infty = o_{\mathbb{P}}(1)$.

Proof of Lemma F.1. For notational convenience, we write $\widehat{\boldsymbol{\beta}}^\lambda_{\mathrm{Lasso}}(\mathcal{D}_j)$ simply as $\widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j)$. Decompose $\overline{\boldsymbol{\beta}}^d - \boldsymbol{\beta}^*$ as

$$\begin{aligned}
\overline{\boldsymbol{\beta}}^d - \boldsymbol{\beta}^* &= \frac{1}{k}\sum_{j=1}^{k}\Big(\widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j) - \boldsymbol{\beta}^* + \frac{1}{n_k} M^{(j)} X^{(j)T} X^{(j)} \big(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j)\big)\Big) \\
&\qquad + \frac{1}{k}\sum_{j=1}^{k} \frac{1}{n_k} M^{(j)} X^{(j)T} \varepsilon^{(j)} \\
&= \frac{1}{k}\sum_{j=1}^{k} \big(I - M^{(j)}\widehat{\Sigma}^{(j)}\big)\big(\widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j) - \boldsymbol{\beta}^*\big) + \frac{1}{k}\sum_{j=1}^{k}\frac{1}{n_k} M^{(j)} X^{(j)T}\varepsilon^{(j)},
\end{aligned}$$

hence $\sqrt{n}(\overline{\boldsymbol{\beta}}^d - \boldsymbol{\beta}^*) = \boldsymbol{Z} + \boldsymbol{\Delta}$, where

$$\boldsymbol{Z} = \frac{1}{\sqrt{k}}\sum_{j=1}^{k}\frac{1}{\sqrt{n_k}} M^{(j)} X^{(j)T}\varepsilon^{(j)} \text{ and } \boldsymbol{\Delta} = \frac{\sqrt{n}}{k}\sum_{j=1}^{k}\big(I - M^{(j)}\widehat{\Sigma}^{(j)}\big)\big(\widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j) - \boldsymbol{\beta}^*\big).$$

Defining $\boldsymbol{\Delta}^{(j)} = \big(I - M^{(j)}\widehat{\Sigma}^{(j)}\big)\big(\widehat{\boldsymbol{\beta}}^{\lambda}(\mathcal{D}_j) - \boldsymbol{\beta}^*\big)$, we have

$$\|\boldsymbol{\Delta}^{(j)}\|_\infty \le \|\boldsymbol{\Delta}^{(j)}\|_1 \le \|M^{(j)}\widehat{\Sigma}^{(j)} - I\|_{\max}\|\widehat{\boldsymbol{\beta}}^{\lambda}(\mathcal{D}_j) - \boldsymbol{\beta}^*\|_1$$

by Hölder's inequality, where $\|I - M^{(j)}\widehat{\Sigma}^{(j)}\|_{\max} \le \vartheta_1$ by the definition of $M^{(j)}$ and, for $\lambda = C\sigma^2\sqrt{\log d/n_k}$,

$$\mathbb{P}\left(\big\|\widehat{\boldsymbol{\beta}}^{\lambda}(\mathcal{D}_j) - \boldsymbol{\beta}^*\big\|_1^2 > C\frac{s^2\log(2d)}{n_k} + t\right) \le \exp\left(-\frac{cn_k t}{s^2\sigma^2}\right) \tag{F.1}$$

by Bühlmann and van de Geer (2011). We thus bound the expectation of the ℓ_1 loss by

$$\begin{aligned}\mathbb{E}\left[\big\|\widehat{\boldsymbol{\beta}}^{\lambda}(\mathcal{D}_j) - \boldsymbol{\beta}^*\big\|_1^2\right] &\le \frac{2Cs^2\log(2d)}{n_k} + \int_0^\infty \exp\left(-\frac{cn_k t}{s^2\sigma^2}\right)dt \\ &\le \frac{2Cs^2\log(2d)}{n_k} + \frac{s^2\sigma^2}{cn_k}.\end{aligned} \tag{F.2}$$

Define the event $\mathcal{E}^{(j)} := \big\{\big\|\widehat{\boldsymbol{\beta}}^{\lambda}(\mathcal{D}_j) - \boldsymbol{\beta}^*\big\|_1 \le s\sqrt{C\log(2d)/n_k}\big\}$ for $j = 1,\dots,k$. $\|\boldsymbol{\Delta}^{(j)}\|_\infty \le \Delta_1^{(j)} + \Delta_2^{(j)} + \Delta_3^{(j)}$ where

$$\begin{aligned}\Delta_1^{(j)} &= \|M^{(j)}\widehat{\Sigma}^{(j)} - I\|_{\max}\|\widehat{\boldsymbol{\beta}}^{\lambda}(\mathcal{D}_j) - \boldsymbol{\beta}^*\|_1 \mathbb{1}\{\mathcal{E}^{(j)}\} \\ &\quad - \mathbb{E}\big[\|M^{(j)}\widehat{\Sigma}^{(j)} - I\|_{\max}\|\widehat{\boldsymbol{\beta}}^{\lambda}(\mathcal{D}_j) - \boldsymbol{\beta}^*\|_1 \mathbb{1}\{\mathcal{E}^{(j)}\}\big] \\ \Delta_2^{(j)} &= \|M^{(j)}\widehat{\Sigma}^{(j)} - I\|_{\max}\|\widehat{\boldsymbol{\beta}}^{\lambda}(\mathcal{D}_j) - \boldsymbol{\beta}^*\|_1 \mathbb{1}\{\mathcal{E}^{(j)c}\} \\ &\quad - \mathbb{E}[\|M^{(j)}\widehat{\Sigma}^{(j)} - I\|_{\max}\|\widehat{\boldsymbol{\beta}}^{\lambda}(\mathcal{D}_j) - \boldsymbol{\beta}^*\|_1 \mathbb{1}\{\mathcal{E}^{(j)c}\}] \quad \text{and} \\ \Delta_3^{(j)} &= \mathbb{E}[\|M^{(j)}\widehat{\Sigma}^{(j)} - I\|_{\max}\|\widehat{\boldsymbol{\beta}}^{\lambda}(\mathcal{D}_j) - \boldsymbol{\beta}^*\|_1].\end{aligned}$$

Consider $\Delta_1^{(j)}$, $\Delta_2^{(j)}$ and $\Delta_3^{(j)}$ in turn. By Hoeffding's inequality, we have for any $t > 0$,

$$\mathbb{P}\left(\frac{1}{k}\sum_{j=1}^{k}\Delta_1^{(j)} > t\right) \le \exp\left(-\frac{n_k k t^2}{Cs^2\vartheta_1^2\log(2d)}\right) \le \exp\left(-\frac{n_k n t^2}{Cs^2\log^2(2d)}\right). \tag{F.3}$$

By Markov's inequality,

$$\begin{aligned}
\mathbb{P}\left(\frac{1}{k}\sum_{j=1}^{k}\Delta_2^{(j)} > t\right) &\le \frac{\sum_{j=1}^{k}\mathbb{E}[\Delta_2^{(j)}]}{kt} \\
&\le 2t^{-1}\mathbb{E}\big[\|M^{(j)}\widehat{\Sigma}^{(j)} - I\|_{\max}\|\widehat{\boldsymbol{\beta}}^{\lambda}(\mathcal{D}_j) - \boldsymbol{\beta}^*\|_1\,\mathbb{1}\{\mathcal{E}^{(j)c}\}\big] \tag{F.4} \\
&\le 2t^{-1}\vartheta_1\sqrt{\mathbb{E}\big[\|\widehat{\boldsymbol{\beta}}^{\lambda}(\mathcal{D}_j) - \boldsymbol{\beta}^*\|_1^2\big]\mathbb{P}(\mathcal{E}^{(j)c})} \\
&\le Ct^{-1}\sqrt{\frac{\log d}{n_k}\cdot\frac{s^2\log(2d)}{n_k}d^{-c}} \le Ct^{-1}sn_k^{-1}d^{-c/2}\log d, \tag{F.5}
\end{aligned}$$

where the penultimate inequality follows from Jensen's inequality. Finally, by Jensen's inequality again,

$$\begin{aligned}
\frac{1}{k}\sum_{j=1}^{k}\Delta_3^{(j)} &= \mathbb{E}[\|M^{(j)}\widehat{\Sigma}^{(j)} - I\|_{\max}\|\widehat{\boldsymbol{\beta}}^{\lambda}(\mathcal{D}_j) - \boldsymbol{\beta}^*\|_1] \\
&\le \vartheta_1\sqrt{\mathbb{E}\left[\|\widehat{\boldsymbol{\beta}}^{\lambda}(\mathcal{D}_j) - \boldsymbol{\beta}^*\|_1^2\right]} \le C\frac{s\log d}{n_k}. \tag{F.6}
\end{aligned}$$

Combining (F.3), (F.5) and (F.6),

$$\begin{aligned}
\mathbb{P}\left(\|\boldsymbol{\Delta}\|_\infty > 3C\sqrt{n}\cdot\frac{s\log d}{n_k}\right) &\le \sum_{u=1}^{3}\mathbb{P}\left(\frac{1}{k}\sum_{j=1}^{k}\Delta^{(j)} > C\sqrt{n}\cdot\frac{s\log d}{n_k}\right) \\
&\le \exp(-ckn) + d^{-c/2} \to 0, \tag{F.7}
\end{aligned}$$

and taking $k = o\big((s\log d)^{-1}\sqrt{n}\big)$ delivers $\|\boldsymbol{\Delta}\|_\infty = o_{\mathbb{P}}(1)$. □

PROOF OF THEOREM A.1.

$$\begin{aligned}
\overline{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}} &= \frac{1}{k}\sum_{j=1}^{k}((X^{(j)})^T X^{(j)})^{-1}(X^{(j)})^T\boldsymbol{Y}^{(j)} - (X^TX)^{-1}X^T\boldsymbol{Y} \\
&= \frac{1}{k}\sum_{j=1}^{k}\left(\left(X^{(j)T}X^{(j)}/n_k\right)^{-1} - (X^TX/n)^{-1}\right)X^{(j)T}\boldsymbol{\varepsilon}^{(j)}/n_k \tag{F.8} \\
&= \frac{1}{k}\sum_{j=1}^{k}\left(\left(X^{(j)T}X^{(j)}/n_k\right)^{-1} - \Sigma^{-1}\right)X^{(j)T}\boldsymbol{\varepsilon}^{(j)}/n_k \\
&\quad + \big(\Sigma^{-1} - (X^TX/n)^{-1}\big)\,X^T\boldsymbol{\varepsilon}/n.
\end{aligned}$$

For simplicity, denote $X^{(j)T}X^{(j)}/n_k$ by $S_X^{(j)}$, X^TX/n by S_X, $(S_X^{(j)})^{-1}-(\Sigma)^{-1}$ by $D_1^{(j)}$ and $(\Sigma)^{-1}-{S_X}^{-1}$ by D_2. For any $\tau\in\mathbb{R}$, define an event $\mathcal{E}^{(j)}=\{\|(S_X^{(j)})^{-1}\|_2\le 2/C_{\min}\}\cap\{\|S_X^{(j)}-\Sigma\|_2\le(\delta_1\vee\delta_1^2)\}$ for all $j=1,\dots,k$, where $\delta_1=C_1\sqrt{d/n_k}+\tau/\sqrt{n_k}$, and an event $\mathcal{E}=\{\|(S_X)^{-1}\|_2\le 2/C_{\min}\}\cap\{\|S_X-\Sigma\|_2<(\delta_2\vee\delta_2^2)\}$, where $\delta_2=C_1\sqrt{d/n}+\tau/\sqrt{n}$. Note that by Lemma F.2 and F.5, the probability of both $(\mathcal{E}^{(j)})^c$ and $\mathcal{E}^c$ are very small. In particular

$$\mathbb{P}(\mathcal{E}^c)\le\exp(-cn)+\exp(-c_1\tau^2)\text{ and }\mathbb{P}((\mathcal{E}^{(j)})^c)\le\exp(-cn/k)+\exp(-c_1\tau^2).$$

Then, letting $\mathcal{E}_0:=\bigcap_{j=1}^{k}\mathcal{E}^{(j)}$, an application of the union bound and Lemma F.9 delivers

$$\begin{aligned}
\mathbb{P}\left(\|\overline{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}\|_2>t\right)&\le\mathbb{P}\left(\left\{\Big\|\frac{1}{k}\sum_{j=1}^{k}(X^{(j)}D_1^{(j)})^T\boldsymbol{\varepsilon}^{(j)}/n_k\Big\|_2>t/2\right\}\cap\mathcal{E}_0\right)\\
&+\mathbb{P}\left(\left\{\|(XD_2)^T\boldsymbol{\varepsilon}/n\|_2>t/2\right\}\cap\mathcal{E}\right)+\mathbb{P}(\mathcal{E}_0^c)+\mathbb{P}(\mathcal{E}^c)\\
&\le 2\exp\left(d\log(6)-\frac{t^2C_{\min}^3 n}{32C_3\sigma_1^2\delta_1^2}\right)+k\exp(-cn/k)+(k+1)\exp(-c_1\tau^2).
\end{aligned}$$

When $d\to\infty$ and $\log n=o(d)$, choose $\tau=\sqrt{d/c_1}$ and $\delta_1=O(\sqrt{kd/n})$. Then there exists a constant C such that

$$\mathbb{P}\left(\|\overline{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}\|_2>C\frac{\sqrt{k}d}{n}\right)\le(k+3)\exp(-d)+k\exp(-\frac{cn}{k}).$$

Otherwise choose $\tau=\sqrt{\log n/c_1}$ and $\delta_1=O(\sqrt{k\log n/n})$. Then there exists a constant C such that

$$\mathbb{P}\left(\|\overline{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}\|_2>C\frac{\sqrt{k}\log n}{n}\right)\le\frac{k+3}{n}+k\exp(-\frac{cn}{k}).$$

Overall, we have

$$\mathbb{P}\left(\|\overline{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}\|_2>C\frac{\sqrt{k}(d\vee\log n)}{n}\right)\le ck\exp(-(d\vee\log n))+k\exp(-cn/k),$$

which leads to the final conclusion. □

Proof of Corollary A.3. Define an event

$$\mathcal{E}=\{\|\overline{\boldsymbol{\beta}}^d-\boldsymbol{\beta}^*\|_\infty\le 2C\sqrt{\log d/n}\},$$

then by the condition on the minimal signal strength and Lemma 4.1, for some constant C' we have

$$\begin{aligned}
&\mathbb{P}\left(\|\overline{\boldsymbol{\beta}}^r - \widehat{\boldsymbol{\beta}}^o\|_2 > C'\frac{\sqrt{k}(s \vee \log n)}{n}\right)\\
&\quad\le \mathbb{P}\left(\left\{\|\overline{\boldsymbol{\beta}}^r - \widehat{\boldsymbol{\beta}}^o\|_2 > C'\frac{\sqrt{k}(s \vee \log n)}{n}\right\} \cap \mathcal{E}\right) + \mathbb{P}(\mathcal{E}^c)\\
&\quad\le \mathbb{P}\left(\left\{\|\overline{\boldsymbol{\beta}}^o - \widehat{\boldsymbol{\beta}}^o\|_2 > C'\frac{\sqrt{k}(s \vee \log n)}{n}\right\} \cap \mathcal{E}\right) + c/d\\
&\quad\le ck\exp(-(s \vee \log n)) + k\exp(-cn/k) + c/d.
\end{aligned}$$

where $\overline{\boldsymbol{\beta}}^o = \frac{1}{k}\sum_{j=1}^{k}(X_S^{(j)T}X_S^{(j)})^{-1}X_S^{(j)T}\boldsymbol{Y}^{(j)}$, which is the average of the oracle estimators on the subsamples. Then the conclusion can be easily validated. □

Proof of Theorem B.1. The following notation is used throughout the proof.

$$\begin{aligned}
&S(\boldsymbol{\beta}) := \nabla^2\ell_n(\boldsymbol{\beta}) = \frac{1}{n}X^T D(X\boldsymbol{\beta})X, \quad S_X^{(j)} := \frac{1}{n_k}X^{(j)T}X^{(j)}\\
&S^{(j)}(\boldsymbol{\beta} := \nabla^2\ell_{n_k}^{(j)}(\boldsymbol{\beta}) = \frac{1}{n_k}X^{(j)T}D(X^{(j)}\boldsymbol{\beta})X^{(j)}, \quad S_X := \frac{1}{n}X^T X.
\end{aligned}$$

For any $j = 1, \ldots, k$, $\widehat{\boldsymbol{\beta}}^{(j)}$ satisfies

$$\nabla\ell_{n_k}^{(j)}(\widehat{\boldsymbol{\beta}}^{(j)}) = \frac{1}{n_k}X^{(j)T}(\boldsymbol{Y}^{(j)} - \boldsymbol{\mu}(X^{(j)}\widehat{\boldsymbol{\beta}}^{(j)})) = 0.$$

Through a Taylor expansion of the left hand side at the point $\boldsymbol{\beta} = \boldsymbol{\beta}^*$, we have

$$\frac{1}{n_k}X^{(j)T}(\boldsymbol{Y}^{(j)} - \boldsymbol{\mu}(X^{(j)}\boldsymbol{\beta}^*)) - S^{(j)}(\widehat{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}^*) - \mathbf{r}^{(j)} = 0,$$

where the remainder term $\mathbf{r}^{(j)}$ is a d dimensional vector with g^{th} component

$$\begin{aligned}
r_g^{(j)} &= \frac{1}{6n_k}(\widehat{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}^*)^T\nabla_{\boldsymbol{\beta}}^2[(\boldsymbol{X}_g^{(j)})^T\boldsymbol{\mu}(X^{(j)}\boldsymbol{\beta})](\widehat{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}^*)\\
&= \frac{1}{6n_k}(\widehat{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}^*)^T X^{(j)T}\mathrm{diag}\{\boldsymbol{X}_g^{(j)} \circ \boldsymbol{\mu}''((X^{(j)}\widetilde{\boldsymbol{\beta}}^{(j)}))\}X^{(j)}(\widehat{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}^*),
\end{aligned}$$

where $\widetilde{\boldsymbol{\beta}}^{(j)}$ is in a line segment between $\widehat{\boldsymbol{\beta}}^{(j)}$ and $\boldsymbol{\beta}^*$. It therefore follows that

$$\widehat{\boldsymbol{\beta}}^{(j)} = \boldsymbol{\beta}^* + (S^{(j)})^{-1}[X^{(j)T}(\boldsymbol{Y}^{(j)} - \boldsymbol{\mu}(X^{(j)}\boldsymbol{\beta}^*)) + n_k\mathbf{r}^{(j)}].$$

A similar equation holds for the global MLE $\widehat{\boldsymbol{\beta}}$:

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}^* + S^{-1}[X^T(\boldsymbol{Y} - \boldsymbol{\mu}(X\boldsymbol{\beta}^*)) + n\mathbf{r}],$$

where for $g = 1, \ldots, d$,

$$r_g = \frac{1}{6n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^T X^T \mathrm{diag}\{\boldsymbol{X}_g \circ \boldsymbol{\mu}''((X\widetilde{\boldsymbol{\beta}}^{(j)}))\} X(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*).$$

Therefore we have

$$\begin{aligned}\frac{1}{k}\sum_{j=1}^{k}\widehat{\boldsymbol{\beta}}^{(j)} - \widehat{\boldsymbol{\beta}} &= \frac{1}{k}\sum_{j=1}^{k}\left\{(S^{(j)})^{-1} - \Sigma^{-1}\right\} X^{(j)T}(\boldsymbol{Y}^{(j)} - \boldsymbol{\mu}(X^{(j)}\boldsymbol{\beta}^*)) \\ &- \left\{S^{-1} - \Sigma^{-1}\right\} X^T(\boldsymbol{Y} - \boldsymbol{\mu}(X\boldsymbol{\beta}^*)) + \boldsymbol{R} = \boldsymbol{B} + \boldsymbol{R},\end{aligned}$$

where $\boldsymbol{R} = (1/k)\sum_{j=1}^{k}(S^{(j)})^{-1}\mathbf{r}^{(j)} - S^{-1}\mathbf{r}$. We next derive stochastic bounds for $\|\boldsymbol{B}\|_2$ and $\|\boldsymbol{R}\|_2$ respectively, but to study the appropriate threshold, we introduce the following events with probability that approaches one under appropriate scaling. For $j = 1, \ldots, k$ and $\kappa, \tau, t > 0$,

$$\begin{aligned}\mathcal{E}^{(j)} &:= \{\|(S^{(j)})^{-1}\|_2 \le 2/C_{\min}\} \cap \Big\{\frac{\|S^{(j)} - \Sigma\|_2}{\delta_1 \vee \delta_1^2} \le 1\Big\} \cap \{\|S_X^{(j)}\|_2 \le 2C_{\max}\}, \\ \mathcal{E} &:= \{\|S^{-1}\|_2 \le 2/L_{\min}\} \cap \big\{\|S - \Sigma\|_2 \le (\delta_2 \vee \delta_2^2)\big\} \cap \{\|S_X\|_2 \le 2C_{\max}\}, \\ \mathcal{F}^{(j)} &:= \left\{\|\widehat{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}^*\|_2 > t\right\}, \quad \mathcal{F} := \left\{\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 > t\right\},\end{aligned}$$

where $\delta_1 = C_1\sqrt{d/n_k} + \tau/\sqrt{n_k}$ and $\delta_2 = C_1\sqrt{d/n_k} + \tau/\sqrt{n}$. Denote the intersection of all the above events by $\mathcal{A}$. Note that Condition 3.6 implies that $\sqrt{b''(\boldsymbol{X}_i^T\boldsymbol{\beta})}\boldsymbol{X}_i$ are i.i.d. sub-Gaussian vectors, so by Lemmas F.2, F.5, F.4 and F.11, we have

$$\begin{aligned}\mathbb{P}(\mathcal{A}^c) &\le (2k+1)\exp\left(-\frac{cn}{k}\right) + (k+1)\exp(-c_1\tau^2) \\ &+ 2k\exp\left(d\log 6 - \frac{nC_{\min}^2 L_{\min}^2 t^2}{2^{11}C_{\max}U_2\phi k}\right)\end{aligned}$$

We first consider the bounded design, i.e., Condition 3.6 (ii). In order to bound $\|\boldsymbol{R}\|_2$, we first derive an upper bound for $r_g^{(j)}$. Under the event $\mathcal{A}$, by Lemma E.5 we have

$$\max_{1\le g\le d, 1\le j\le k} r_g^{(j)} \le \frac{1}{3}MU_3C_{\max}t^2 \text{ and } \max_{1\le g\le d} r_g \le \frac{1}{3}MU_3C_{\max}t^2.$$

It follows that, under $\mathcal{A}$,

$$\|\boldsymbol{R}\|_2 \le \frac{2}{3} M\sqrt{d}\, U_3 C_{\max} t^2. \tag{F.9}$$

Note that $\boldsymbol{B}$ is very similar to the RHS of Equation (F.8). Now we use essentially the same proof strategy as in the OLS part to bound $\|\boldsymbol{B}\|_2$. Following similar notations as in OLS, we denote $(S^{(j)})^{-1} - \Sigma^{-1}$ by $D_1^{(j)}$, $S^{-1} - \Sigma^{-1}$ by D_2, $\boldsymbol{Y}^{(j)} - \boldsymbol{\mu}(X^{(j)}\boldsymbol{\beta}^*)$ by $\boldsymbol{\varepsilon}^{(j)}$ and $\boldsymbol{Y} - \boldsymbol{\mu}(X\boldsymbol{\beta}^*)$ by $\boldsymbol{\varepsilon}$. For concision, we relegate the details of the proof to Lemma F.10, which delivers the following stochastic bound on $\|\boldsymbol{B}\|_2$.

$$\mathbb{P}(\{\|\boldsymbol{B}\|_2 > t_1\} \cap \mathcal{A}) \le 2\exp\left(d\log(6) - \frac{C_{\min}^4 L_{\min}^2 n t_1^2}{128\phi U_2 C_{\max}(\delta_1 \vee \delta_1^2)^2}\right). \tag{F.10}$$

Combining Equation (F.10) with (F.9) leads us to the following inequality.

$$\begin{aligned}
\mathbb{P}&\left(\|\overline{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}\|_2 > \frac{2}{3} M\sqrt{d}\, U_3 C_{\max} t^2 + t_1\right) \le (2k+1)\exp\left(-\frac{cn}{k}\right) \\
&+ (k+1)\exp(-c_1\tau^2) + (k+1)\exp\left(d\log 6 - \frac{C_{\min}^2 L_{\min}^2 n t^2}{2^{11} C_{\max} U_2 \phi k}\right) \\
&+ 2\exp\left(d\log 6 - \frac{C_{\min}^4 L_{\min}^2 n t_1^2}{128\phi U_2 C_{\max}(\delta_1 \vee \delta_1^2)^2}\right).
\end{aligned}$$

Choose $t = t_1 = \sqrt{d/n_k}$ and, when $d \gg \log n$, choose $\tau = \sqrt{d/c_1}$ and $\delta_1 = O(\sqrt{kd/n})$. Then there exists a constant $C > 0$ such that

$$\mathbb{P}\left(\|\overline{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}\|_2 > C\frac{kd^{3/2}}{n}\right) \le (2k+1)\exp(-\frac{cn}{k}) + 2(k+2)\exp(-d).$$

When it is not true that $d \gg \log n$, choose $\tau = \sqrt{\log n/c_1}$ and $\delta = O(\sqrt{k\log n/n})$. Then there exists a constant $C > 0$ such that

$$\mathbb{P}\left(\|\overline{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}\|_2 > C\frac{k\sqrt{d}\log n}{n}\right) \le (2k+1)\exp(-\frac{cn}{k}) + \frac{k+3}{n}.$$

Overall, we have

$$\mathbb{P}\left(\|\overline{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}\|_2 > C\frac{k\sqrt{d}(d \vee \log n)}{n}\right) \le ck\exp(-cn/k) + ck\exp(-c(d \vee \log n)),$$

which leads to the final conclusion. □

Proof of Corollary B.3. Define an event

$$\mathcal{E} = \{\|\overline{\boldsymbol{\beta}}^d - \boldsymbol{\beta}^*\|_\infty \le 2C\sqrt{\log d/n}\},$$

then by the conditions of Corollary B.3 and results of Lemma 4.6 and Theorem B.1,

$$\begin{aligned}
&\mathbb{P}\left(\|\overline{\boldsymbol{\beta}}^r - \widehat{\boldsymbol{\beta}}^o\|_2 > C'\frac{k\sqrt{s}(s \vee \log n)}{n}\right)\\
&\quad\le \mathbb{P}\left(\left\{\|\overline{\boldsymbol{\beta}}^r - \widehat{\boldsymbol{\beta}}^o\|_2 > C'\frac{k\sqrt{s}(s \vee \log n)}{n}\right\} \cap \mathcal{E}\right) + \mathbb{P}(\mathcal{E}^c)\\
&\quad\le \mathbb{P}\left(\left\{\|\overline{\boldsymbol{\beta}}^o - \widehat{\boldsymbol{\beta}}^o\|_2 > C'\frac{k\sqrt{s}(s \vee \log n)}{n}\right\} \cap \mathcal{E}\right) + c/d\\
&\quad\le ck\exp(-(s \vee \log n)) + k\exp(-cn/k) + c/d.
\end{aligned}$$

where $\overline{\boldsymbol{\beta}}^o = \frac{1}{k}\sum_{j=1}^{k} \widehat{\boldsymbol{\beta}}^o(\mathcal{D}_j)$, $\widehat{\boldsymbol{\beta}}^o(\mathcal{D}_j) = \mathrm{argmax}_{\boldsymbol{\beta}\in\mathbb{R}^d, \boldsymbol{\beta}_{S^c}=0}\, \ell^{(j)}(\boldsymbol{\beta})$ and C' is a constant. Then it is not hard to see that the final conclusion is true. □

Proof of Lemma 4.1. According to Lemma F.1, we have $\sqrt{n}(\overline{\boldsymbol{\beta}}^d - \boldsymbol{\beta}^*) = \boldsymbol{Z} + \boldsymbol{\Delta}$, where $\boldsymbol{Z} = \frac{1}{\sqrt{k}}\sum_{j=1}^k \frac{1}{\sqrt{n_k}} M^{(j)} X^{(j)T}\boldsymbol{\varepsilon}^{(j)}$. In (F.7), we prove that $\|\boldsymbol{\Delta}\|_\infty/\sqrt{n} \le Csk\log d/n$ with probability larger than $1 - \exp(-ckn) - d^{-c/2} \ge 1 - c_1/d$ for some constant c_1. Since $\widehat{\boldsymbol{\beta}}^d$ is a special case of $\overline{\boldsymbol{\beta}}^d$ when $k = 1$, we also have $\sqrt{n}(\widehat{\boldsymbol{\beta}}^d - \boldsymbol{\beta}^*) = \boldsymbol{Z} + \boldsymbol{\Delta}_1$, where (F.7) gives $\|\boldsymbol{\Delta}\|_\infty/\sqrt{n} \le Cs\log d/n$. Therefore, we have $\|\overline{\boldsymbol{\beta}}^d - \widehat{\boldsymbol{\beta}}^d\|_\infty \le Csk\log d/n$ with high probability.

It only remains to bound the rate of $\|\boldsymbol{Z}\|_\infty/\sqrt{n}$. By Condition 3.2, conditioning on $\{\boldsymbol{X}_i\}_{i=1}^n$, we have for any $\ell = 1, \ldots, d$,

$$\begin{aligned}
\mathbb{P}\Big(|Z_\ell|/\sqrt{n} > t \,\Big|\, \{\boldsymbol{X}_i\}_{i=1}^n\Big) &= \mathbb{P}\Big(\Big|\frac{1}{n}\sum_{j=1}^k \boldsymbol{M}_\ell^{(j)T} X^{(j)T}\boldsymbol{\varepsilon}^{(j)}\Big| > t \,\Big|\, \{\boldsymbol{X}_i\}_{i=1}^n\Big)\\
&\le 2\exp\Big(-\frac{cnt^2}{\kappa^2 Q_\ell}\Big),
\end{aligned}\tag{F.11}$$

where κ is the variance proxy of ε defined in Condition 3.2 and

$$Q_\ell = \frac{1}{n}\sum_{j=1}^k \|X^{(j)}\boldsymbol{M}_\ell^{(j)T}\|_2^2.$$

Let $Q_{\max} = \max_{1\le\ell\le d} Q_\ell$. Applying the union bound to (F.11), we have

$$\mathbb{P}\Big(\|\boldsymbol{Z}\|_\infty/\sqrt{n} > t \,\Big|\, \{\boldsymbol{X}_i\}_{i=1}^n\Big) \le \mathbb{P}\Big(\max_{1\le\ell\le d} |Z_\ell|/\sqrt{n} > t \,\Big|\, \{\boldsymbol{X}_i\}_{i=1}^n\Big)$$
$$\le \sum_{\ell=1}^d \mathbb{P}\Big(|Z_\ell|/\sqrt{n} > t \,\Big|\, \{\boldsymbol{X}_i\}_{i=1}^n\Big) \le 2d\exp\Big(-\frac{cnt^2}{\kappa^2 Q_{\max}}\Big).$$

Let $t = \sqrt{2\kappa^2 Q_{\max}\log d/(cn)}$, then with conditional probability $1-2/d$,

$$\|\boldsymbol{Z}\|_\infty/\sqrt{n} \le \sqrt{\kappa^2 Q_{\max}\log d/(cn)}. \tag{F.12}$$

The last step is to bound $Q_{\max}$. By the definition of Q_ℓ, we have

$$\begin{aligned} Q_\ell &= \frac{1}{k}\sum_{j=1}^k \boldsymbol{M}_\ell^{(j)T}\widehat{\Sigma}^{(j)}\boldsymbol{M}_\ell^{(j)} \le \frac{1}{k}\sum_{j=1}^k [\boldsymbol{\Omega}]_\ell^T\widehat{\Sigma}^{(j)}[\boldsymbol{\Omega}]_\ell \\ &= \frac{1}{k}\sum_{j=1}^k \frac{1}{n_k}\sum_{i\in\mathcal{D}_j}(\boldsymbol{X}_i^T[\boldsymbol{\Omega}]_\ell)^2 = \frac{1}{n}\sum_{i=1}^n(\boldsymbol{X}_i^T[\boldsymbol{\Omega}]_\ell)^2, \end{aligned} \tag{F.13}$$

where $\Omega = \Sigma^{-1}$. The inequality is due to the fact that $M_\ell^{(j)}$ is the minimizer in (3.4). By condition (3.2) and the connection between sub-Gaussian and subexponential distributions, the random variable $(\boldsymbol{X}_i^T\boldsymbol{\Omega}_\ell)^2$ satisfies

$$\sup_{q\ge1} q^{-1}\big(\mathbb{E}|(\boldsymbol{X}_i^T\boldsymbol{\Omega}_\ell)^2|^q\big)^{1/q} \le 4\kappa^2\Omega_{\ell\ell}.$$

Therefore, by Bernstein's inequality for subexponential random variables, we have

$$\mathbb{P}\Big(\Big|\frac{1}{n}\sum_{i=1}^n(\boldsymbol{X}_i^T[\boldsymbol{\Omega}]_\ell)^2 - \mathbb{E}[\boldsymbol{X}_1^T[\boldsymbol{\Omega}]_\ell]^2\Big| > t\Big) \le 2\exp\Big(-c\Big(\frac{nt^2}{16\kappa^4\Omega_{\ell\ell}^2}\Big)\wedge\Big(\frac{nt}{4\kappa^2\Omega_{\ell\ell}}\Big)\Big).$$

Applying the union bound again, we have

$$\mathbb{P}\Big(\max_{1\le\ell\le d}\Big|\frac{1}{n}\sum_{i=1}^n(\boldsymbol{X}_i^T[\boldsymbol{\Omega}]_\ell)^2 - \mathbb{E}[\boldsymbol{X}_1^T[\boldsymbol{\Omega}]_\ell]^2\Big| > 8\kappa^2\Omega_{\ell\ell}\sqrt{\frac{\log d}{cn}}\Big)$$
$$\le \sum_{j=1}^d \mathbb{P}\Big(\Big|\frac{1}{n}\sum_{i=1}^n(\boldsymbol{X}_i^T[\boldsymbol{\Omega}]_\ell)^2 - \mathbb{E}[\boldsymbol{X}_1^T[\boldsymbol{\Omega}]_\ell]^2\Big| > 8\kappa^2\Omega_{\ell\ell}\sqrt{\frac{\log d}{cn}}\Big) \le 2/d.$$

Therefore, with probability $1 - 2/d$, there exist a constant C_1 such that

$$Q_{\max} = \max_{1 \le \ell \le d} Q_\ell \le \max_{1 \le \ell \le d} \left| \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{X}_i^T \boldsymbol{\Omega}_\ell)^2 - \mathbb{E}[\boldsymbol{X}_1^T \boldsymbol{\Omega}_\ell]^2 \right| + \mathbb{E}[\boldsymbol{X}_1^T \boldsymbol{\Omega}_\ell]^2$$

$$\le 8\kappa^2 \boldsymbol{\Omega}_{jj} \sqrt{\frac{\log d}{cn}} + \Omega_{jj} \le C_1,$$

where the last inequality is due to Condition 3.1. By (F.12), we have with probability $1 - 4/d$, $\|\boldsymbol{Z}\|_\infty / \sqrt{n} \le \sqrt{\kappa^2 C_1 \log d/(cn)}$. Combining this with the result on $\|\boldsymbol{\Delta}\|_\infty$ delivers the rate in the lemma. □

Proof of Lemma 4.6. The strategy of proving this lemma is similar to the proof of Lemma 4.1. In the proof of Lemma E.7 and Theorem 3.8, we have shown that

$$(\overline{\boldsymbol{\beta}}^d - \boldsymbol{\beta}^*) = \underbrace{-\frac{1}{k} \sum_{j=1}^{k} \widehat{\Theta}^{(j)T} \nabla \ell_{n_k}^{(j)}(\boldsymbol{\beta}^*)}_{\mathbf{T}} + \frac{1}{k} \sum_{j=1}^{k} \boldsymbol{\Delta}_j,$$

where the remainder term for each j is

$$\boldsymbol{\Delta}_j = \left(I - \widehat{\Theta}^{(j)T} \frac{1}{n_k} \sum_{i \in \mathcal{I}_j} b''(\widetilde{\eta}_i) \boldsymbol{X}_i \boldsymbol{X}_i^T \right) (\widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j) - \boldsymbol{\beta}^*)$$

and $\widetilde{\eta}_i = t\boldsymbol{X}_i^T \boldsymbol{\beta}^* + (1-t)\boldsymbol{X}_i^T \widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j)$ for some $t \in (0,1)$. We bound $\boldsymbol{\Delta}_j$ by decomposing it into three terms:

$$\begin{aligned}
\|\boldsymbol{\Delta}_j\|_\infty \le\ & \underbrace{\Big\| \big(I - \Theta^* \frac{1}{n_k} \sum_{i \in \mathcal{I}_j} b''(\boldsymbol{X}_i^T \boldsymbol{\beta}^*) \boldsymbol{X}_i \boldsymbol{X}_i^T \big) (\widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j) - \boldsymbol{\beta}^*) \Big\|_\infty}_{I_1} \\
& + \underbrace{\Big\| \Theta^* \frac{1}{n_k} \sum_{i \in \mathcal{I}_j} (b''(\boldsymbol{X}_i^T \widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j)) - b''(\boldsymbol{X}_i^T \boldsymbol{\beta}^*)) \boldsymbol{X}_i \boldsymbol{X}_i^T \Big) (\widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j) - \boldsymbol{\beta}^*) \Big\|_\infty}_{I_2} \\
& + \underbrace{\Big\| (\widehat{\Theta}^{(j)} - \Theta^*)^T \frac{1}{n_k} \sum_{i \in \mathcal{I}_j} b''(\boldsymbol{X}_i^T \widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j)) \boldsymbol{X}_i \boldsymbol{X}_i^T \Big) (\widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j) - \boldsymbol{\beta}^*) \Big\|_\infty}_{I_3}.
\end{aligned}$$

By Hoeffding's inequality and Condition 3.3, the first term is bounded by

$$|I_1| \le \Big\| I - \Theta^* \frac{1}{n_k} \sum_{i \in \mathcal{I}_j} b''(\boldsymbol{X}_i^T \boldsymbol{\beta}^*) \boldsymbol{X}_i \boldsymbol{X}_i^T \Big\|_{\max} \Big\| \widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j) - \boldsymbol{\beta}^* \Big\|_1 \le C \frac{sk \log d}{n}, \tag{F.14}$$

with probability $1-c/d$. By Condition 3.6 (iii), Condition 3.7 (iv) and Lemma E.4, we have with probability $1-c/d$,

$$|I_2| \le \max_i \big\|\boldsymbol{\Theta}^*\boldsymbol{X}_i\big\|_\infty \frac{1}{n_k}\sum_{i\in\mathcal{I}_j} U_3[\boldsymbol{X}_i(\widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j)-\boldsymbol{\beta}^*)]^2 \le C\frac{sk\log d}{n}. \tag{F.15}$$

Finally, we bound I_3 by with probability $1-c/d$,

$$|I_3| \le \frac{\sqrt{U_2}}{n_k}\sqrt{\sum_{i\in\mathcal{I}_j} b''(\boldsymbol{X}_i^T\widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j))[\boldsymbol{X}_i^T(\widehat{\Theta}^{(j)}-\Theta^*)]^2}\sqrt{\sum_{i\in\mathcal{I}_j}[\boldsymbol{X}_i(\widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j)-\boldsymbol{\beta}^*)]^2} \tag{F.16}$$
$$\le C\frac{(s_1\vee s)k\log d}{n},$$

where the last inequality is due to Lemma E.4 and Lemma C.4 of Ning and Liu (2014).

Combining (F.14) - (F.16) and applying the union bound, we have

$$\Big\|\frac{1}{k}\sum_{j=1}^k \boldsymbol{\Delta}_j\Big\|_\infty \le \max_j \|\boldsymbol{\Delta}_j\|_\infty = O_P\Big(\frac{(s_1\vee s)k\log d}{n}\Big).$$

Therefore, we only need to bound the infinity norm of the leading term $\mathbf{T}$. By Condition 3.7 and equation (E.3), we have with probability $1-c/d$,

$$\begin{aligned}&\max_{1\le j\le k}\max_{1\le v\le d}\|\widehat{\boldsymbol{\Theta}}_v^{(j)}-\boldsymbol{\Theta}_v^*\|_1 \le Cs_1\sqrt{\log d/n} \quad \text{and}\\ &\Big\|\frac{1}{k}\sum_{j=1}^k \nabla\ell_{n_k}^{(j)}(\boldsymbol{\beta}^*)\Big\|_\infty \le C\sqrt{\log d/n}.\end{aligned} \tag{F.17}$$

This, together with Condition 3.6 and Condition 3.7 give the bound,

$$\|\mathbf{T}\|_\infty \le \Big(M\max_{v,j}\|\widehat{\boldsymbol{\Theta}}_v^{(j)}-\boldsymbol{\Theta}_v^*\|_1 + \max_i\|\boldsymbol{X}_i^T\boldsymbol{\Theta}^*\|_\infty\Big)\Big\|\frac{1}{k}\sum_{j=1}^k \nabla\ell_{n_k}^{(j)}(\boldsymbol{\beta}^*)\Big\|_\infty$$
$$\le C\Big(\sqrt{\frac{\log d}{n}}+\frac{s_1\log d}{n}\Big),$$

with probability $1-c/d$. Since $\widehat{\boldsymbol{\beta}}^d$ is a special case of $\overline{\boldsymbol{\beta}}^d$ when $k=1$, the proof of the lemma is complete. □

We borrow the following two lemmas on concentration inequalities from Vershynin (2010) for the proof later.

Lemma F.2. Suppose X is a $n\times d$ matrix that has independent sub-Gaussian rows $\{\boldsymbol{X}_i\}_{i=1}^n$. Denote $\mathbb{E}(\boldsymbol{X}_i\boldsymbol{X}_i^T)$ by Σ, then we have

$$\mathbb{P}\left(\|\frac{1}{n}X^TX-\Sigma_X\|_2\geq(\delta\vee\delta^2)\right)\leq\exp(-c_1t^2),$$

where $t\geq 0$, $\delta=C_1\sqrt{d/n}+t/\sqrt{n}$ and C_1 and c_1 are both constants depending only on $\|\boldsymbol{X}_i\|_{\psi_2}$.

Proof. See Theorem 5.39 and Remark 5.40 in Vershynin (2010). □

Lemma F.3. (Bernstein-type inequality) Let $X_1,\dots,X_n$ be independent centered sub-exponential random variables, and $M=\max_{1\leq i\leq n}\|X_i\|_{\psi_1}$. Then for every $a=(a_1,\dots,a_n)\in\mathbb{R}^n$ and every $t\geq 0$, we have

$$\mathbb{P}\left(\sum_{i=1}^n a_iX_i\geq t\right)\leq\exp\left(-C_2\min\left(\frac{t^2}{M^2\|a\|_2^2},\frac{t}{M\|a\|_\infty}\right)\right).$$

Proof. See Proposition 5.16 in Vershynin (2010). □

Lemma F.4. Suppose X is a $n\times d$ matrix that has independent sub-gaussian rows $\{\mathbf{x}_i\}_{i=1}^n$. If $\lambda_{\max}(\Sigma)\leq C_{\max}$ and $d\ll n$, then for all $M>C_{\max}$, there exists a constant $c>0$ such that when n and d are sufficiently large,

$$\mathbb{P}\left(\left\|\frac{1}{n}X^TX\right\|_2\geq M\right)\leq\exp(-cn).$$

Proof. Apply Lemma F.2 with $t=\sqrt{cn/c_1}$, where $(\sqrt{c/c_1}\vee c/c_1)<M-C_{\max}$, and it follows that

$$\mathbb{P}\left(\left\|\frac{1}{n}X^TX-\Sigma\right\|_2\geq(\delta\vee\delta^2)\right)\leq\exp(-cn).$$

Since $d\ll n$, we obtain $(\delta\vee\delta^2)\to\sqrt{c/c_1}$, which completes the proof. □

Lemma F.5. Suppose X is a $n\times d$ matrix that has independent sub-Gaussian rows $\{\boldsymbol{X}_i\}_{i=1}^n$. $\mathbb{E}\boldsymbol{X}_i=\mathbf{0}$, $\lambda_{\min}(\Sigma)\geq C_{\min}>0$ and $d\ll n$. For all $m<C_{\min}$, there exists a constant $c>0$ such that when n and d are sufficiently large,

$$\mathbb{P}\left(\left\|\left(\frac{1}{n}X^TX\right)^{-1}\right\|_2\geq\frac{1}{m}\right)=\mathbb{P}\left(\lambda_{\min}\left(\frac{1}{n}X^TX\right)\leq m\right)\leq\exp(-cn).$$

Proof. It is easy to check the following inequality. For any two symmetric and semi-definite $d \times d$ matrices A and B, we have

$$\lambda_{\min}(A) \geq \lambda_{\min}(B) - \|A - B\|_2 \ ,$$

because for any vector $\mathbf{x}$ satisfying $\|\mathbf{x}\|_2 = 1$, we have

$$\|A\boldsymbol{x}\|_2 = \|B\boldsymbol{x} + (A-B)\boldsymbol{x}\|_2 \geq \|B\boldsymbol{x}\|_2 - \|(A-B)\boldsymbol{x}\|_2 \geq \lambda_{\min}(B) - \|A - B\|_2.$$

Then it follows that

$$\begin{aligned} \mathbb{P}\left(\left\|\left(\frac{1}{n}X^T X\right)^{-1}\right\|_2 \geq \frac{1}{m}\right) &= \mathbb{P}\left(\lambda_{\min}\left(\frac{1}{n}X^T X\right) \leq m\right) \\ &\leq \mathbb{P}\left(\left\|\frac{1}{n}X^T X - \Sigma_X\right\|_2 \geq C_{\min} - m\right) \leq \exp(-cn), \end{aligned}$$

where c satisfies $(\sqrt{c/c_1} \vee c/c_1) < C_{\min} - m$ and the last inequality is an application of Lemma F.2 with $t = \sqrt{cn/c_1}$. □

Lemma F.6. (Hoeffding-type Inequality). Let $X_1,\ldots,X_n$ be independent centered sub-Gaussian random variables, and let $K = \max_i \|X_i\|_{\psi_2}$. Then for every $a = (a_1, \ldots, a_n) \in \mathbb{R}^n$ and every $t > 0$, we have

$$\mathbb{P}\left(\left|\sum_{i=1}^n a_i X_i\right| \geq t\right) \leq e \cdot \exp\left(-\frac{ct^2}{K^2\|a\|_2^2}\right).$$

Lemma F.7. (Sub-exponential is sub-Gaussian squared). A random variable X is a sub-Gaussian if and only if X^2 is sub-exponential. Moreover,

$$\|X\|_{\psi_2}^2 \leq \|X^2\|_{\psi_1} \leq 2\|X\|_{\psi_2}^2.$$

Proof. See Lemma 5.14 in Vershynin (2010). □

Lemma F.8. Let $X_1,\ldots,X_n$ be independent centered sub-Gaussian random variables. Let $\kappa = \max_i \|X_i\|_{\psi_2}$ and $\sigma^2 = \max_i \mathbb{E}X_i^2$. Suppose $\sigma^2 > 1$, then we have

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n X_i^2 > 2\sigma^2\right) \leq \exp\left(-C_2\frac{\sigma^2 n}{\kappa^2}\right).$$

Proof. Combining Lemma F.3 and Lemma F.7 yields the result. □

Lemma F.9. Following the same notation as in the beginning of Proof of Theorem A.1,

$$\mathbb{P}\left(\left\{\|\frac{1}{k}\sum_{j=1}^{k}\frac{(X^{(j)}D_1^{(j)})^T\boldsymbol{\varepsilon}^{(j)}}{n_k}\|_2 > \frac{t}{2}\right\}\cap\mathcal{E}_0\right) \le 6^d\exp\left(-\frac{t^2C_{\min}^3 n}{32C_3s_1^2(\delta_1\vee\delta_1^2)^2}\right)$$

and

$$\mathbb{P}\left(\{\|(XD_2)^T\boldsymbol{\varepsilon}/n\|_2 > t/2\}\cap\mathcal{E}\right) \le \exp\left(d\log(6) - \frac{t^2C_{\min}^3 n}{32C_3s_1^2(\delta_2\vee\delta_2^2)^2}\right).$$

Proof.

(F.18)

$$\mathbb{E}\left(\exp\left(\lambda(D_1^{(j)}\mathbf{v})^T(X^{(j)T}\boldsymbol{\varepsilon}^{(j)}/n_k)\right) \mid X^{(j)}\right)$$
$$= \prod_{i=1}^{n_k}\mathbb{E}\left(\exp((\frac{\lambda\boldsymbol{X}_i^{(j)}}{n_k})^T(D^{(j)}\mathbf{v})\varepsilon_i) \mid X^{(j)}\right) \le \exp\left(C_3\lambda^2s_1^2\sum_{i=1}^{n}(\frac{A_i^{(j)}}{n_k})^2\right),$$

(F.19)

$$\mathbb{E}\left(\exp\left(\lambda(D_2\mathbf{v})^T(X^T\boldsymbol{\varepsilon}/n)\right) \mid X\right) = \prod_{i=1}^{N}\mathbb{E}\left(\exp\left((\lambda\boldsymbol{X}_i/N)^T(D_2\mathbf{v})\varepsilon_i\right) \mid X\right)$$
$$\le \exp\left(C_3\lambda^2s_1^2\sum_{i=1}^{N}A_i^2/n^2\right),$$

where we write $A_i^{(j)}$ and A_i in place of $(\boldsymbol{X}_i^{(j)})^TD_1^{(j)}\mathbf{v}$ and $(\boldsymbol{X}_i)^TD_2\mathbf{v}$ respectively C_3 is an absolute constant, and the last inequality holds because ε_i are sub-Gaussian. Next we provide an upper bound on $\sum_{i=1}^{n_k}(A_i^{(j)})^2$ and $\sum_{i=1}^{n}A_i^2$. Note that

$$\begin{aligned}\sum_{i=1}^{n}(A_i^{(j)})^2 &= \mathbf{v}^TD_1^{(j)}X^TXD_1^{(j)}\mathbf{v}\\ &= \mathbf{v}^T((S_X^{(j)})^{-1}-(\Sigma)^{-1})n_kS_X^{(j)}((S_X^{(j)})^{-1}-(\Sigma)^{-1})\mathbf{v}\\ &= n_k\mathbf{v}^T\Sigma^{-1}(\Sigma-S_X^{(j)})(S_X^{(j)})^{-1}(\Sigma-S_X^{(j)})\Sigma^{-1}\mathbf{v},\end{aligned}$$

and similarly,

$$\sum_{i=1}^{n}A_i^2 = n\mathbf{v}^T\Sigma^{-1}(\Sigma-S_X)(S_X)^{-1}(\Sigma-S_X)\Sigma^{-1}\mathbf{v}.$$

For any $\tau \in \mathbb{R}$, define the event $\mathcal{E}^{(j)} = \{\|(S_X^{(j)})^{-1}\|_2 \le 2/C_{\min}\} \cap \{\|S_X^{(j)} - \Sigma\|_2 \le (\delta_1 \vee \delta_1^2)\}$ for all $j = 1, \dots, k$, where $\delta_1 = C_1\sqrt{d/n_k} + \tau/\sqrt{n_k}$, and the event $\mathcal{E} = \{\|(S_X)^{-1}\|_2 \le 2/C_{\min}\} \cap \{\|S_X - \Sigma\|_2 < (\delta_2 \vee \delta_2^2)\}$, where $\delta_2 = C_1\sqrt{d/n} + \tau/\sqrt{n}$. On $\mathcal{E}^{(j)}$ and $\mathcal{E}$, we have respectively

$$\sum_{i=1}^{n_k} (A_i^{(j)})^2 \le \frac{2n_k}{C_{\min}^3}(\delta_1 \vee \delta_1^2)^2 \text{ and } \sum_{i=1}^{n} A_i^2 \le \frac{2n}{C_{\min}^3}(\delta_2 \vee \delta_2^2)^2.$$

Therefore from Equation (F.18) and (F.19) we obtain

$$\mathbb{E}\left(\exp(\lambda(D_1^{(j)}\mathbf{v})^T(X^{(j)T}\boldsymbol{\varepsilon}^{(j)}/n_k))\,\mathbb{1}\{\mathcal{E}^{(j)}\}\right) \le \exp\left(\frac{2C_3\lambda^2 s_1^2}{C_{\min}^3 n_k}(\delta_1 \vee \delta_1^2)^2\right)$$

and

$$\mathbb{E}\left(\exp(\lambda(D_2\mathbf{v})^T(X^T\boldsymbol{\varepsilon}/n))\,\mathbb{1}\{\mathcal{E}\}\right) \le \exp\left(\frac{2C_3\lambda^2 s_1^2}{C_{\min}^3 N}(\delta_2 \vee \delta_2^2)^2\right).$$

In addition, according to Lemma F.2 and F.5, the probability of both $(\mathcal{E}^{(j)})^c$ and $\mathcal{E}^c$ are very small. More specifically,

$$\mathbb{P}(\mathcal{E}^c) \le \exp(-cn) + \exp(-c_1\tau^2) \text{ and } \mathbb{P}((\mathcal{E}^{(j)})^c) \le \exp(-cn/k) + \exp(-c_1\tau^2).$$

Let $\mathcal{E}_0 := \bigcap_{j=1}^{k} \mathcal{E}^{(j)}$. An application of the Chernoff bound trick leads us to the following inequality.

$$\begin{aligned}
&\mathbb{P}\left(\left\{\frac{1}{k}\sum_{j=1}^{k}(D_1^{(j)}\mathbf{v})^T(X^{(j)T}\boldsymbol{\varepsilon}^{(j)})/n_k > t/2\right\} \cap \mathcal{E}_0\right) \\
&\le \exp(-\lambda t/2)\prod_{j=1}^{k}\mathbb{E}\left(\exp\left(\frac{\lambda}{k}(D_1^{(j)}\boldsymbol{v})^T(X^{(j)T}\boldsymbol{\varepsilon}^{(j)}/n_k)\right)\mathbb{1}\{\mathcal{E}^{(j)}\}\right) \\
&\le \exp\left(-\lambda t/2 + \frac{2C_3\lambda^2 s_1^2}{C_{\min}^3 n}(\delta_1 \vee \delta_1^2)^2\right).
\end{aligned}$$

Minimize the right hand side by λ, then we have

$$\mathbb{P}\left(\left\{\frac{1}{k}\sum_{j=1}^{k}\frac{(D_1^{(j)}\boldsymbol{v})^T(X^{(j)T}\boldsymbol{\varepsilon}^{(j)})}{n_k} > \frac{t}{2}\right\} \cap \mathcal{E}_0\right) \le \exp\left(-\frac{t^2 C_{\min}^3 n}{32C_3 s_1^2(\delta_1 \vee \delta_1^2)^2}\right).$$

Consider the $1/2-$net of $\mathbb{R}^p$, denoted by $\mathcal{N}(1/2)$. Again it is known that $|\mathcal{N}(1/2)| < 6^p$. Using the maximal inequality, we have

$$\begin{aligned}
&\mathbb{P}\left(\left\{\|\frac{1}{k}\sum_{j=1}^{k}(X^{(j)}D_1^{(j)})^T\boldsymbol{\varepsilon}^{(j)}/n_k\big\|_2 > t/2\right\}\cap\mathcal{E}_0\right)\\
&= \sup_{\|\mathbf{v}\|_2=1}\mathbb{P}\left(\left\{\frac{1}{k}\sum_{j=1}^{k}(D_1^{(j)}\boldsymbol{v})^T(X^{(j)T}\boldsymbol{\varepsilon}^{(j)})/n_k > t/2\right\}\cap\mathcal{E}_0\right)\\
&\le \sup_{\mathbf{v}\in\mathcal{N}(1/2)}\mathbb{P}\left(\left\{\frac{1}{k}\sum_{j=1}^{k}(D_1^{(j)}\boldsymbol{v})^T(X^{(j)T}\boldsymbol{\varepsilon}^{(j)})/n_k > t/4\right\}\cap\mathcal{E}_0\right)\\
&\le \exp\left(d\log(6) - \frac{t^2C_{\min}^3 n}{32C_3s_1^2(\delta_1\vee\delta_1^2)^2}\right).
\end{aligned}$$

Proceeding in an analogous fashion, we obtain

$$\mathbb{P}\left(\left\{\|(XD_2)^T\boldsymbol{\varepsilon}/n\|_2 > t/2\right\}\cap\mathcal{E}\right) \le \exp\left(d\log(6) - \frac{t^2C_{\min}^3 n}{32C_3s_1^2(\delta_2\vee\delta_2^2)^2}\right).$$

□

Lemma F.10. Following the same notation as in the proof of Theorem B.1,

$$\mathbb{P}(\{\|\boldsymbol{B}\|_2 > t_1\}\cap\mathcal{A}) \le 2\exp\left(d\log(6) - \frac{C_{\min}^4L_{\min}^2nt_1^2}{128\phi U_2C_{\max}(\delta_1\vee\delta_1^2)^2}\right).$$

Proof. By Lemma E.2, for any $\lambda\in\mathbb{R}$ and $\mathbf{v}$ such that $\|\mathbf{v}\|_2 = 1$, we have

$$\begin{aligned}
&\mathbb{E}\left(\exp(\lambda(D_1^{(j)}\mathbf{v})^T(X^{(j)T}\boldsymbol{\varepsilon}^{(j)}/n_k)) \mid X^{(j)}\right)\\
&= \prod_{i=1}^{n_k}\mathbb{E}\left(\exp((\lambda\boldsymbol{X}_i^{(j)}n_k)^T(D^{(j)}\mathbf{v})\varepsilon_i) \mid X^{(j)}\right) \le \exp\left(\phi U\lambda^2\sum_{i=1}^{n_k}(A_i^{(j)})^2/n_k^2\right)
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}\left(\exp(\lambda(D_2\mathbf{v})^T(X^T\boldsymbol{\varepsilon}/n)) \mid X\right) &= \prod_{i=1}^{n}\mathbb{E}\left(\exp((\lambda\boldsymbol{X}_i/n)^T(D_2\mathbf{v})\varepsilon_i) \mid X\right)\\
&\le \exp\left(\phi U\lambda^2\sum_{i=1}^{n}A_i^2/n^2\right),
\end{aligned}$$

where we write $A_i^{(j)}$ and A_i in place of $(X_i^{(j)})^T D_1^{(j)}\mathbf{v}$ and $(X_i)^T D_2\mathbf{v}$ respectively. Next we give a upper bound on $\sum_{i=1}^{n_k}(A_i^{(j)})^2$ and $\sum_{i=1}^{n} A_i^2$. Note that

$$\begin{aligned}\sum_{i=1}^{n_k}(A_i^{(j)})^2 &= \mathbf{v}^T D_1^{(j)} X^T X D_1^{(j)}\mathbf{v}\\ &= \mathbf{v}^T((S^{(j)})^{-1} - \Sigma^{-1})nS_X((S^{(j)})^{-1} - \Sigma^{-1})\mathbf{v}\\ &= n\mathbf{v}^T\Sigma^{-1}(\Sigma - S^{(j)})(S^{(j)})^{-1}S_X^{(j)}(S^{(j)})^{-1}(\Sigma - S^{(j)})\Sigma^{-1}\mathbf{v}.\end{aligned}$$

Similarly,

$$\sum_{i=1}^{n} A_i^2 = n\mathbf{v}^T\Sigma^{-1}(\Sigma - S)S^{-1}S_X S^{-1}(\Sigma - S)\Sigma^{-1}\mathbf{v}.$$

On $\mathcal{E}^{(j)}$ and $\mathcal{E}$, we have respectively

$$\sum_{i=1}^{n_k}(A_i^{(j)})^2 \le \frac{8C_{\max}n_k}{C_{\min}^4 L_{\min}^2}(\delta_1 \vee \delta_1^2)^2 \text{ and } \sum_{i=1}^{n} A_i^2 \le \frac{8C_{\max}n}{C_{\min}^4 L_{\min}^2}(\delta_2 \vee \delta_2^2)^2.$$

Then it follows that

$$\mathbb{E}\left(\exp(\lambda(D_1^{(j)}\boldsymbol{v})^T(X^{(j)T}\boldsymbol{\varepsilon}^{(j)}/n_k))\,\mathbb{1}\{\mathcal{E}^{(j)}\}\right) \le \exp\left(\frac{8\phi U C_{\max}\lambda^2}{C_{\min}^4 L_{\min}^2 n_k}(\delta_1 \vee \delta_1^2)^2\right)$$

and

$$\mathbb{E}\left(\exp(\lambda(D_2\mathbf{v})^T(X^T\boldsymbol{\varepsilon}/n))\,\mathbb{1}\{\mathcal{E}\}\right) \le \exp\left(\frac{8\phi U C_{\max}\lambda^2}{C_{\min}^4 L_{\min}^2 n}(\delta_2 \vee \delta_2^2)^2\right).$$

Now we follow exactly the same steps as in the OLS part. Denote $\cap_{j=1}^k \mathcal{E}_j$ by $\mathcal{E}_0$. An application of the Chernoff bound technique and the maximal inequality leads us to the following inequality.

$$\begin{aligned}&\mathbb{P}\left(\left\{\|\frac{1}{k}\sum_{j=1}^{k}(X^{(j)}D_1^{(j)})^T\boldsymbol{\varepsilon}^{(j)}/n_k\|_2 > t/2\right\}\cap\mathcal{E}_0\right)\\ &\quad\le \exp\left(d\log(6) - \frac{C_{\min}^4 L_{\min}^2 n t^2}{128\phi U_2 C_{\max}(\delta_1 \vee \delta_1^2)^2}\right).\end{aligned}$$

and

$$\mathbb{P}\left(\left\{\|(XD_2)^T\boldsymbol{\varepsilon}/n\|_2 > t/2\right\}\cap\mathcal{E}\right) \le \exp\left(d\log(6) - \frac{C_{\min}^4 L_{\min}^2 n t^2}{128\phi U_2 C_{\max}(\delta_2 \vee \delta_2^2)^2}\right).$$

We have thus derived an upper bound for $\|\boldsymbol{B}\|_2$ that holds with high probability. Specifically,

$$\mathbb{P}(\{\|\boldsymbol{B}\|_2 > t_1\} \cap \mathcal{A}) \leq \mathbb{P}\left(\left\{\|\frac{1}{k}\sum_{j=1}^{k}(X^{(j)}D_1^{(j)})^T\boldsymbol{\varepsilon}^{(j)}/n_k\|_2 > \frac{t_1}{2}\right\} \cap \mathcal{E}_0\right)$$
$$+ \mathbb{P}\left(\left\{\|\frac{(XD_2)^T\boldsymbol{\varepsilon}}{n}\|_2 > \frac{t_1}{2}\right\} \cap \mathcal{E}\right) \leq 2\cdot 6^d \exp\left(-\frac{C_{\min}^4 L_{\min}^2 n t_1^2}{128\phi U_2 C_{\max}(\delta_1 \vee \delta_1^2)^2}\right).$$

□

Lemma F.11. Under Condition 3.6, for $\tau \leq L_{\min}/(8MC_{\max}U_3\sqrt{d})$ and sufficiently large n and d we have

$$\mathbb{P}(\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 > \tau) \leq \exp\left(d\log 6 - \frac{nC_{\min}^2 L_{\min}^2 \tau^2}{2^{11}C_{\max}U_2\phi}\right) + 2\exp(-cn).$$

Proof. The notation is that introduced in the proof of Theorem B.1. We further define $\Sigma(\boldsymbol{\beta}) := \mathbb{E}(b''(X^T\boldsymbol{\beta})XX^T)$ as well as the event $\mathcal{H} := \{\ell_n(\boldsymbol{\beta}^*) > \max_{\boldsymbol{\beta}\in\partial\mathcal{B}_\tau}\ell_n(\boldsymbol{\beta})\}$, where $\mathcal{B}_\tau = \{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \leq \tau\}$. Note that as long as the event $\mathcal{H}$ holds, the MLE falls in $\mathcal{B}_\tau$, therefore the proof strategy involves showing that $\mathbb{P}(\mathcal{H})$ approaches 1 at certain rate. By the Taylor expansion,

$$\ell_n(\boldsymbol{\beta}) - \ell_n(\boldsymbol{\beta}^*) = (\boldsymbol{\beta} - \boldsymbol{\beta}^*)^T\boldsymbol{v} - \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)^T S(\widetilde{\boldsymbol{\beta}})(\boldsymbol{\beta} - \boldsymbol{\beta}^*) = A_1 + A_2,$$

where $S(\boldsymbol{\beta}) = (1/n)X^T D(X\boldsymbol{\beta})X$, $\widetilde{\boldsymbol{\beta}}$ is some vector between $\boldsymbol{\beta}$ and $\boldsymbol{\beta}^*$, $\boldsymbol{v} = (1/n)X^T(\boldsymbol{Y} - \boldsymbol{\mu}(X\boldsymbol{\beta}^*))$, $A_1 = (\boldsymbol{\beta}-\boldsymbol{\beta}^*)^T\boldsymbol{v} - (1/2)(\boldsymbol{\beta}-\boldsymbol{\beta}^*)^T S(\boldsymbol{\beta}^*)(\boldsymbol{\beta}-\boldsymbol{\beta}^*)$ and $A_2 = -(1/2)(\boldsymbol{\beta} - \boldsymbol{\beta}^*)^T(S(\widetilde{\boldsymbol{\beta}}) - S(\boldsymbol{\beta}^*))(\boldsymbol{\beta} - \boldsymbol{\beta}^*)$.

Define the event $\mathcal{E} := \{\lambda_{\min}[S(\boldsymbol{\beta}^*)] \geq L_{\min}/2\}$, where $L_{\min}$ is the same constant in Condition 3.6. Note that by Condition 3.6 (ii), $\sqrt{b''(\boldsymbol{X}_i^T\boldsymbol{\beta})}\boldsymbol{X}_i$ is a sub-Gaussian random vector. Then by Condition 3.6 (iii) and Lemma F.5, for sufficiently large n and d we have $\mathbb{P}(\mathcal{E}^c) \leq \exp(-cn)$. Therefore on the event $\mathcal{E}$,

$$A_1 \leq \tau(\|\boldsymbol{v}\|_2 - \frac{L_{\min}}{4}\tau).$$

We next show that, under an appropriate choice of τ, $|A_2| < L_{\min}\tau^2/8$ with high probability. We first consider Condition 3.6 (ii). Define $\mathcal{F} := \{\|X^T X/n\|_2 \leq 2C_{\max}\}$. By Lemma F.4, we have $\mathbb{P}(\mathcal{F}^c) \leq \exp(-cn)$. By

Lemma E.5, on the event $\mathcal{F}$, we have

$$\begin{aligned} A_2 &\le \max_{1\le i\le n} |b''(\boldsymbol{X}_i^T\widetilde{\boldsymbol{\beta}}) - b''(\boldsymbol{X}_i^T\boldsymbol{\beta}^*)| C_{\max}\tau^2 \\ &\le MU_3\sqrt{d}\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \cdot C_{\max}\tau^2 \\ &\le MC_{\max}U_3\sqrt{d}\tau^3 \le \frac{L_{\min}\tau^2}{8}, \end{aligned}$$

where the last inequality holds if we choose $\tau \le L_{\min}/(8MC_{\max}U_3\sqrt{d})$. Now we obtain the following probabilistic upper bound on $\mathcal{H}^c$, which we later prove to be negligible.

$$\begin{aligned} \mathbb{P}(\mathcal{H}^c) &\le \mathbb{P}(\mathcal{H}^c \cap \mathcal{E} \cap \mathcal{F}) + \mathbb{P}(\mathcal{E}^c) + \mathbb{P}(\mathcal{F}^c) \\ &\le \mathbb{P}\left(\left\{\|\boldsymbol{v}\|_2 \ge \frac{L_{\min}\tau}{8}\right\} \cap \mathcal{E} \cap \mathcal{F}\right) + \mathbb{P}(\mathcal{E}^c) + \mathbb{P}(\mathcal{F}^c). \end{aligned} \tag{F.20}$$

Since each component of $\mathbf{v}$ is a weighted average of i.i.d. random variables, the effect of concentration tends to make $\|\mathbf{v}\|_2$ very small with large probability, which inspires us to study the moment generating function and apply the Chernoff bound technique. By Lemma E.2, for any constant $\mathbf{u} \in \mathbb{R}^d$, $\|\mathbf{u}\|_2 = 1$ and let $a_i = \mathbf{u}^T\boldsymbol{X}_i$, then we have for any $t \in \mathbb{R}$,

$$\begin{aligned} \mathbb{E}\left(\exp(t\langle\mathbf{u},\mathbf{v}\rangle) \mid X\right) &= \prod_{i=1}^n \mathbb{E}\left(\exp\left(\frac{ta_i}{n}(Y_i - \mu(\boldsymbol{X}_i^T\boldsymbol{\beta}))\right) \mid X\right) \\ &\le \exp\left(\frac{\phi U_2 t^2}{2n^2}\sum_{i=1}^n a_i^2\right) \\ &= \exp\left(\frac{\phi U_2 t^2}{2n} \cdot \frac{\mathbf{u}^T X^T X \mathbf{u}}{n}\right). \end{aligned}$$

It follows that

$$\mathbb{E}\exp(t\langle\mathbf{u},\mathbf{v}\rangle\, \mathbb{1}\{\mathcal{E}\cap\mathcal{F}\}) \le \exp\left(\frac{\phi C_{\max}U_2t^2}{2n}\right).$$

By the Chernoff bound technique, we obtain

$$\mathbb{P}(\{\langle\mathbf{u},\mathbf{v}\rangle > \varepsilon\} \cap \mathcal{E} \cap \mathcal{F}) \le \exp\left(-\frac{n\varepsilon^2}{8C_{\max}U_2\phi}\right).$$

Consider a $1/2-$net of $\mathbb{R}^d$, denoted by $\mathcal{N}(1/2)$. Since

$$\|\mathbf{v}\|_2 = \max_{\|\mathbf{u}\|_2=1}\langle\mathbf{u},\mathbf{v}\rangle \le 2\max_{\mathbf{u}\in\mathcal{N}(1/2)}\langle\mathbf{u},\mathbf{v}\rangle,$$

it follows that

$$\begin{aligned}\mathbb{P}(\{\|\mathbf{v}\|_2 > \frac{L_{\min}\tau}{8}\} \cap \mathcal{E} \cap \mathcal{F}) &\leq \mathbb{P}\left(\left\{\max_{\mathbf{u}\in\mathcal{N}(1/2)} \langle \mathbf{u}, \mathbf{v}\rangle > \frac{L_{\min}\tau}{16}\right\} \cap \mathcal{E} \cap \mathcal{F}\right) \\ &\leq 6^d \exp\left(-\frac{nL_{\min}^2\tau^2}{2^{10}\phi C_{\max}U_2}\right) \\ &= \exp\left(d\log 6 - \frac{nC_{\min}^2 L_{\min}^2\tau^2}{2^{11}C_{\max}U_2\phi}\right).\end{aligned}$$

Finally combining the result above with Equation (F.20) delivers the conclusion. □

Remark F.12. Simple calculation shows that when $d = o(\sqrt{n})$, $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 = O_{\mathbb{P}}(\sqrt{d/n})$. When d is a fixed constant, $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 = O_{\mathbb{P}}(\sqrt{1/n})$.

Department of Operations Research
and Financial Engineering
Princeton University
Princeton, New Jersey 08544
USA
E-mail: h.battey@imperial.ac.uk
jqfan@princeton.edu
hanliu@princeton.edu
junweil@princeton.edu
ziweiz@princeton.edu

Department of Mathematics
Imperial College London
London, SW7 2AZ
UK
E-mail: h.battey@imperial.ac.uk

School of Data Science
Fudan University
Shanghai, China
E-mail: jqfan@princeton.edu